DOE NETL

Wellbore Data Quality & Availability in the U.S.

MLEF Summer Research Report

Megan Lynn Tucker

8-3-2019

This research was supported in part by the National Energy Technology Laboratory (NETL), sponsored by the U.S. Department of Energy (DOE) Office of Fossil Energy (FE) Mickey Leland Energy Fellowship (MLEF) program, and administered by the Oak Ridge Institute for Science and Education (ORISE).

Table of Contents

INTRODUCTION

Objective

This project's objective was to study publicly available wellbore datasets from the following states of interest: Alaska, California, Colorado, Kansas, North Dakota, Oklahoma, Pennsylvania, and Texas. These states were chosen for their historically high oil and gas production rates. The regulatory agencies distributing the datasets were assessed for their data availability—was there an online database, downloadable dataset, or online viewer—and ease of navigating the website. The fields and contents of the downloadable datasets were analyzed. Each field was catalogued by name, the acronyms and abbreviations deciphered using the metadata, and the categorical, continuous, and binary variables identified for a more in-depth data quality comparison across databases. The contents of each field were cleaned, removing dummy values standing in for no data. The count and percent complete of each field was recorded. Lastly, each state's fields were compared using these questions: (1) what fields did they or did they not have in common (2) how full were those fields and (3) which states did not contain information on pertinent fields?

Findings from each state were synthesized via preliminary statistical analysis to provide a foundational geographic understanding of wellbore data availability. Graphs and maps were constructed to better visualize the data's spread and relationships. These findings will accelerate database construction, visualization, and data discovery, processing, analysis, and communication by informing machine learning tools and applications. The datasets may be difficult to parse through as they are complicated and non-standardized. Currently, humans sift through the data to verify this usefulness, so automating this process would allow researchers to focus on data analysis instead of data collection.

Background & Motivation

Scientific studies are often limited by data fidelity; it is crucial to understand the inherent biases, variability, and uncertainties in raw data when interpreting results. Awareness of data inconsistencies can provide insight into regulation shifts or influential policy changes, specifically related to oil and gas. Wellbore integrity trends can be better understood through studying these historic events in conjunction with specific wellbore attributes like cement type, casing diameter, spud date, and production history. Analyzing state oil and gas datasets will provide awareness of the physical qualities of wellbores. More importantly it will highlight the quality and availability of data about each states' wellbores. Fundamentally, wellbore data is the foundation for understanding wellbore patterns, statistically and geographically. Studying wellbore data availability, ease of access, database structures, and regulatory bodies can shed light on and evaluate data deficiencies in states of interest and other locations. As wellbore data is used for modeling

various situations involving oil and gas wells, it is important to know about biases within the data and how it could potentially skew the model.

Oil and gas extraction are regulated at the state level, including exploring and developing oil and gas wells, distance between wells and property lines, waste prevention, health and safety, taxation, and data collection (1). Therefore, there is little consistency in the data reported by each state; what and how much information must be reported and what is publicly available or not determined state by state. This makes commercial databases like Drilling Info and Information Handling Services (IHS) appealing—they process wellbore data into a digestible and uniform format. Both databases use information supplied by state regulatory agencies, suggesting they would be the same. However, the numbers of records and fields differs. Evaluating the root sources can provide insight into how much the data are manipulated in these proprietary databases. This may also reveal what uncertainties are introduced through processing.

CARBON STORAGE

Modeling carbon storage is one application of wellbore data. Carbon storage is a topic of interest because it is a method for decreasing CO_2 in the atmosphere while we transition from predominantly using fossil fuels to renewables (1). Eliminating reliance on fossil fuels could drastically decrease the global carbon output. However, fossil fuels have many advantages like ease of transport and storage, availability, low cost, and quantity (2). Finding a way to decrease CO_2 output while allowing the continued use of fossil fuels is ideal. Carbon capture and storage is endorsed by both the International Energy Agency (IEA), IEA Greenhouse Gas R&D Program (IEAGHG), and the U.S. EPA as one such method.

The IEA hopes to stabilize the concentration of CO_2 in the atmosphere at 450 parts per million, which is expected to increase the global temperature by 2°C. To accomplish this, an annual 1.4 Gt of CO_2 must be captured and stored by 2030 (3). In theory 8,000 to 15,000 Gt of CO_2 could be stored in the ground (2), as CO_2 can be stored as a supercritical fluid at depths below 800 meters (about 2,600 feet) on most places on Earth (1). Supercritical CO_2 is highly dense, so it does not require much storage space.

There is concern that the acidic environment created by CO_2 mixing with well water could corrode steel and Portland cement lined wellbores (4) (5), leading to leakage. This impacts groundwater quality (6), the environment, human safety, and economic resources like oil and gas (5). Another concern is that pumping CO_2 into the ground could incite manmade seismic activity. Studies have shown that leakage usually occurs at easily monitored, isolated spots like faults, fractures, and wells. This is likely because wells create a pathway through the primary seal (caprock) for CO_2 to migrate to the surface (5).

Because of the potential risks of carbon storage, computational models are being constructed to assess potential danger. Carbon storage models rely on wellbore data to

predict future trends because oil and natural gas reservoirs are ideal locations for carbon storage; they have held crude oil for millions of years, are well studied due to exploration for producing hydrocarbons, and the infrastructure to transport and store CO_2 already exists (2). However, many of the most favorable sites for carbon storage have 100's or 1,000's of wells drilled into them (5).

When evaluating a site for carbon storage, there are several factors to consider: environmental impact, storage capacity, CO_2 retention time, leakage potential, uptake rate, and cost (2). As wells are highly likely to leak, wellbore integrity[1] must also be evaluated. This can be done for individual wells or generalized by studying the general trends of wellbore data (7). Examples of wellbore data include well age, CO_2 presence, licensee, completion interval depth, surface casing depth, total depth, well density, geographic area, well type, oil price and regulatory changes, and abandonment method.

APPROACH

Data

Publicly available datasets from eight states were evaluated. The datasets were retrieved from its respective state regulatory agency. The amount of data available for each state was investigated. California, Colorado, North Dakota, Oklahoma, and Texas were the primarily chosen; Alaska and Kansas were later added to provide more breadth to the study.

Fields of interest include API number[2], spud date[3], completion date, abandoned date, activity code[4], operator name, field name[5], formation name[6], depth, latitude and longitude, production date, well status[7], production data (cumulative oil and gas, monthly oil and gas), and injection data (cumulative gas or other, monthly gas or other). Many of the states have different names for some fields despite their contents being the same. Activity code and well status were difficult to differentiate between; each state seems to have a different

[1] A well's ability to remain isolated from the outside geological system while preventing the vertical movement of fluids (19).

[2] A unique, permanent, numeric identifier assigned to wells oil or gas producing wells. Digits 1-2 are the state code, 3-5 are the county code, 6-10 are a unique number within the county, 11-12 are a unique bottom hole location of the well, and 13-14 are the "well event sequencing code." A well event is a well deepening, recompletion, or re-entry of a plugged wellbore (20) (21).

[3] The first day of drilling.

[4] The IHS dataset has five different activity codes: PERMIT – Well permitted but not spudded; DRILLING IN PROGRESS – Well spudded and is being drilled; COMPLETED WELL - Well completed, but not all data has been received and processed; COMPLETED WELLS / WHCS (well history control system) - All data received, processed, and transferred to the IHS historical file; ABANDONED LOCATION – Proposed well site where no drilling occurred and was abandoned by the operator.

[5] Name of the oil or gas field.

[6] Name of the geologic formation the oil or gas field is located in.

[7] In the IHS dataset this is essentially a more detailed version of the activity code with 321 possible values.

definition for well status and no state had a field labeled "activity code." Another issue was variations of the same type of data, particularly for depth, latitude, and longitude—these three fields can be measured many ways, and often, datasets had multiple measurements.

ALASKA

The Alaska Oil and Gas Conservation Commission hosts Alaska's oil and gas wellbore data. The database is located on the public search engine, Data Miner 3. The database is updated every Monday through Friday at 7:00pm Alaska Time. Results can be exported as an Excel file or CSV. They do not appear to have a viewer (8).

The dataset does not have abandoned date, activity code, formation name, production date, production, or injection data fields. Two separate fields are provided for depth: driller true depth—original recorded depth—and true vertical depth—vertical distance from the wellhead to the base of the well. The dataset also provides four different variations of latitude and longitude for the wellhead and bottom hole each: NAD27, NAD83[8], reported, and calculated. Something interesting to note about the API numbers is that the 6th through 14th digits are almost all entirely zeros, which could be interpreted as null values. It is not uncommon for states to omit the 11th through 14th digits or the 1st and 2nd, so this is rather unique. The county code may have been omitted because Alaska is divided into boroughs instead of counties. This is a reasonable assumption because the boroughs do not cover the entire land area of the state.

CALIFORNIA

The California Department of Conservation hosts California's oil and gas wellbore data. There are several promising datasets, making it difficult to select a single dataset for this project. The dataset generated through using Well Search—an online well database providing oil and gas well monthly production and injection information from 1977 to present—was chosen because it has the most comprehensive data. The dataset can be exported as an Excel file by deselecting Operator (Active Only). GIS Mapping: All Wells, which provides oil and gas well locations and records published by the California Department of Conservation, Division of Oil Gas and Geothermal Resources (DOGGR), was another promising dataset but was ultimately discarded because it did not include completion and abandon dates. Shapefiles and CSVs are also downloadable via the online viewer, Well Finder (9).

The dataset does not have formation name, depth, production date, production data, or injection data fields. The API numbers are 8 digits; they are missing the first two—the state code (05)—and last four digits—the directional sidetrack and event sequence codes. The activity code specifies if the well is active, idle, plugged, unknown, cancelled, or buried-

[8] The North American Datum of 1983 (NAD83) is a geographic reference frame which replaced NAD27.

idle. The pool well type[9] provides information identifying the well type. Extra geographic information is provided like area code, area name, GIS source code, datum code, and base meridian. From 1997 to 2010 California was split up into 6 districts, later revised to 4 in 2011; these are indicated by the section field. Lastly, the dataset identifies if the well is under the jurisdiction of the Bureau of Land Management (BLMWell), if it is a dry hole (DryHole[10]), and if it revived hydraulic stimulation treatment (HydFrac[11]).

COLORADO

The Colorado Oil and Gas Conservation Commission hosts Colorado's oil and gas wellbore data. Like California, Colorado has multiple datasets. The GIS dataset was chosen for this project; this dataset is downloadable as a Shapefile. Other datasets under consideration were the Oil and Gas Well Analytical Data and the Daily Activity Dashboard (Pending Permits). The Oil and Gas Analytical Data is exported as an Access file. It unfortunately lacks many of the pertinent fields for this project, yet it also contains several fields of interest like Sample Type—for example, Bradenhead Test or Drillstem. There is also a viewer which provides detailed information about specific wells. The data found on the viewer cannot be downloaded. All datasets are updated monthly (10).

The dataset does not have completion date, abandoned date, activity code, formation name, well status, production data, or injection data fields. Four separate fields are provided with information about API: API, API_County, API_Seq, and API_Label. The entries in API_Label concatenates the information from the previous fields into 10 digits long API numbers missing the directional sidetrack and event sequence codes, the last four digits of the API number. Like Alaska, two depth measurements are provided: true vertical depth and max measured depth—total length of the wellbore measured along the well path.

KANSAS

The Kansas Corporation Commission hosts Kansas' oil and gas wellbore data and the Kansas Geological Survey hosts the master list of oil and gas wells; this is downloaded as a text file. The online database can be used to search for and download information about specific wells. When the entire state is selected there is no link to save the data. There is also an online Viewer (11).

The dataset does not have formation name, production date, production data, or injection data fields. The KID field contains a temporary, unique ID assigned by the Kansas Geological Survey. Similarly, OIL_KID and GAS_KID are the Kansas Geological Survey

---

[9] Possible well types: AI = Air Injection; OG = Oil & Gas; DG = Dry Gas; PM = Pressure Maintenance; GD = Gas Disposal; SC = Cyclic Steam; GS = Gas Storage; SF = Steam Flood; LG = Liquified Gas; WD = Water Disposal; MW = Ground Monitoring; WF = Water Flood; NW = New; WS = Water Source; OB = Observation.

[10] The well never produced commercial quantities of hydrocarbons.

[11] Also called hydraulic fracturing.

ID codes for the oil and gas production data respectively. The API numbers are mostly 10 digits, missing the last four, though a small number of them do have their last four digits. The name of the original operator and current operator is also provided—the name of the current operator is used for the duration of analysis. The STATUS[12] field provides information of the type of well via abbreviations; STATUS2[13] provides additional status information from the Kansas Corporation Commission.

NORTH DAKOTA

The North Dakota Department of Mineral Resources hosts North Dakota's oil and gas wellbore data. North Dakota provides a variety of free, publicly available resources, including a viewer and online database. Unfortunately, downloading the complete index of all permitted wells in North Dakota is only available via Premium Subscription Services. Because of this, North Dakota was not included in the data (12).

OKLAHOMA

The Oklahoma Corporation Commission hosts Oklahoma's oil and gas wellbore data. There is no viewer, but there is an online database called the Well Data System and an online repository containing various information like injection volumes, monthly oil and gas productions, and orphaned wells. The Basic Well Completion Master File contains well identification and location information for all wells in Oklahoma from 1985 to present. The file is updated monthly and is exported as a CSV (13).

Oklahoma had substantially more data and more fields than the other states—746,552 entries and 104 fields (the other datasets had between 10,000 and 400,000 entries and 20 to 50 fields). The sheer quantity of data is likely why there are so many erroneous entries, like misplaced decimals, negative latitude, positive longitude, or dates before the year

---

[12] STATUS Values: CBM = produced coalbed methane; CBM-P&A = produced coalbed methane, since plugged and abandoned; D&A = never produced, now plugged and abandoned; EOR = enhanced oil recovery well; EOR-P&A = enhanced oil recovery well, since plugged and abandoned; GAS = produced natural gas; GAS-P&A = produced natural gas, since plugged and abandoned; INJ = salt water disposal well or other injection well; INJ-P&A = salt water disposal well or other injection well, since plugged and abandoned; INTENT = proposed well, not yet drilled; LOC = well that was never actually drilled; O&G = produced oil and gas; O&G-P&A = produced oil and gas, since plugged and abandoned; OIL = produced oil; OIL-P&A = produced oil, since plugged and abandoned; OTHER = may not be an energy well, since water research wells and road construction wells are in database under some conditions; OTHER-P&A = miscellaneous well since plugged; SWD = salt water disposal well; SWD-P&A = salt water disposal well, since plugged and abandoned.

[13] STATUS2 Values: Approved for Plugging - CP-1 Received; Approved Intent to Drill; Authorized Injection Well; Cancelled API Number; Converted to EOR Well; Converted to Producing Well; Converted to SWD Well; Expired Intent to Drill (C-1); Inactive Well; Injection Authorization Terminated; Injection Well Split to Another Dkt; KCC Fee Fund Plugging; ON LIST; Plugged and Abandoned; Producing; Recompleted; Spudded; UIC Application Denied; UIC Application Dismissed; UIC Application Withdrawn; Unknown; Unplugged Former Injection Well; Well Drilled.

1900. The values for latitude and longitude were also given in a variety of formats—degrees, meters, feet—meaning the data required a substantial amount of cleaning before analysis.

Of the relevant data, only the field name and the abandoned date are missing. The API numbers are mostly 10 digits; the 14 digits long entries end in 0000. The spud date, completion date, and production date contain many dates listed as 1/1/1900. This is still a feasible date for well spudding and completion so we cannot disregard these values. However, it is unlikely that such a quantity of wells was completed at the same time. The operator name contains several entries of OTC/OCC NOT ASSIGNED; those values were taken to be blank. Like Colorado, measured total depth and true vertical depth are provided. However, many of the cells were assigned 0, which is problematic in the same way as the dates above. Many values in latitude and longitude also contain 0, which were again assumed blank.

PENNSYLVANIA

The Pennsylvania Department of Environmental Protection hosts Pennsylvania's oil and gas wellbore data. As with Colorado and California, it was difficult finding a dataset because of the numerous promising ones: Oil and Gas Well Inventory, Well Formations Report, Oil and Gas Locations, Oil and Gas Locations (Conventional), Oil and Gas Locations (Unconventional). Oil and Gas Locations (Conventional and Unconventional)—two datasets combined into a single set—was eventually chosen. Oil and Gas Well Inventory contained information about other states, Well Formations Report only had 62,263 entries when there should be around 200,000, and Oil and Gas Locations contained fewer fields. Pennsylvania also has a viewer which provides details about specific wells (14).

The dataset does not contain completion date, abandoned date, field name, formation name, depth, production date, production data, or injection data fields. The API numbers are only 8 digits, and are missing the state, directional sidetrack, and event sequence codes. The spud dates all end in T00:00:00.000Z. There are several entries of "1800-01-01" in the spud date field. Unlike 1/1/1900, it is reasonable to assume this value represents no data, as the first recorded oil well was constructed in 1859.

TEXAS

The Texas Railroad Commission hosts Texas' oil and gas wellbore data. Texas has an online database which can search for individual wells, and an online viewer. The viewer can export a CSV of all wells within a radius of 2.5 miles with a maximum of 1,000 wells in the area. However, the full wellbore dataset is not freely available. Those willing to pay have access to records for most of the wells ever drilled in Texas by API number. The dataset supposedly includes completion date, plugging date, formation, and other

information related to wellbores. Because the data were not publicly accessible, Texas was not included in any further analysis (15).

Method

Once the publicly available datasets were retrieved and studied, they were analyzed individually and compared to other states. Like data types were mapped together and the quantity of data in each field was compared. Preliminary statistics were used to indicate which fields and states lacked data. Graphs of the spud date were created to highlight periods of increased drilling and potential dummy variables. Maps were used to show where drilling was most prominent and where the data may be lacking.

DATA CATALOGUING

The count—number of nonempty entries—of each field was determined for each state dataset using R. Depending on the dataset and its fields, blank entries were represented differently, such as empty cells, spaces, or dummy variables like "unknown," "0," or "1/1/1900." Therefore, the data were first cleaned, replacing these dummy values with null values. After finding the count of each field, the percent complete was calculated by dividing the count by the total number of entries.

Metadata for all state datasets except Oklahoma and Pennsylvania was readily available. Using the metadata, like attributes were mapped together, i.e. data types were mapped together based on content, not name. Oklahoma and Pennsylvania were mapped to the other datatypes based on each field's contents. Following this, a spreadsheet of the count and percent of relevant datatypes within each state dataset was created.

Statistical analysis was limited to summary statistics[14] as the data collected was deliberately observed and not found through random sampling. The statistics were represented visually using box charts. The percent of data was compared as opposed to the amount of data because each state has a different quantity of data available to study. The statistics were calculated and graphed using Excel.

GEOSTATISTICAL ANALYSIS

ArcPro was used to generate calendar heat charts and data clocks for the spud dates of each state. The spud date was used because it is the only date field present in all state datasets. The key for each map is slightly different because the quantity of data for each state through time is different. Furthermore, the spread of dates of each state is different, so although two graphs may look similar, they may represent different information.

[14] Summary statistics include information like the mean, median, range, and skew of the data.

The heat chart is a visual mapping of how often certain dates occur—each grid cell represents an intersection between month and day of the month. The count of incidents occurring at such times are represented by different colors (16). This can help determine if a certain date is used as a dummy variable, such as 1/1/1900.

Similarly, data clocks reveal seasonal data over the span of 100 years (17). This can highlight dummy variables, times of peak production, and where data may be incomplete. For instance, a well may only have a year associated with its spud date, so it may be listed as being spudded on January 1. Two data clocks were made for each state whose range of spud dates were greater than 100 years, one pre-1960 and one post-1960. Again, it is important to compare the key as there tends to be a different breakdown in values between the two clocks. Line charts were also constructed to aid in understanding the overarching trends of the dataclocks.

ArcPro was also used to generate maps displaying the spread of wells across each state. The points coordinate to a color gradient from blue to pink depending on the spud date—bright blue represents the older spud date and bright pink represents the more recent spud dates.

RESULTS

Bar charts containing percent complete by state were constructed for each field of interest. The percent complete was visualized by two types of box plots; the first shows percent complete by state for each field, i.e. states are on the horizontal axis and the second shows percent complete of each field by state, i.e. fields are on the horizontal axis.

Further analysis of the datasets focused on spud date because it contains important information and is one of the few fields the six states have in common. The spud dates were analyzed three different ways: calendar heat chart, data clock, and geospatially. The geospatial analysis took the form of studying where wells with similar spud dates were located geographically. Temporal heat maps were constructed showing the well construction timeline.

Statistical

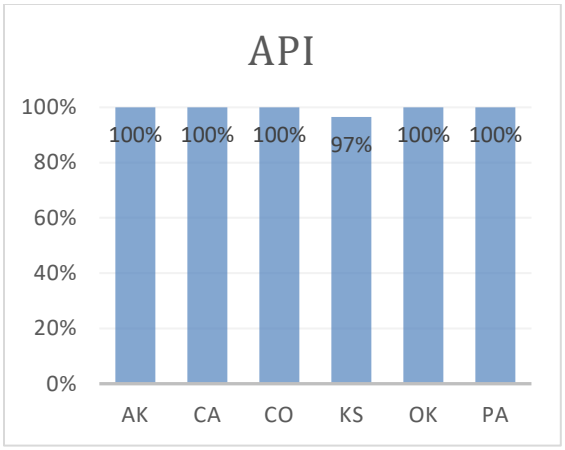
All surveyed states contained information on API, spud date, operator name, latitude, and longitude fields. Blank sections on the graph indicate that there was no corresponding state.

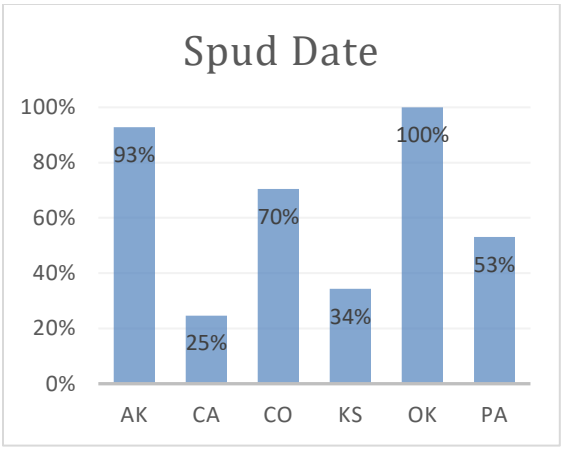
Oklahoma was the only state to provide information on formation name, first production date, production data and injection data. Those charts were omitted; percent complete of each field was 99.60%, 100.00%, 100.00% and 100.00% respectively. The Appendix of this paper contains additional bar charts with the count of each field by state. The fields are labeled as they were in the original dataset, so like data may not be labeled the same way.

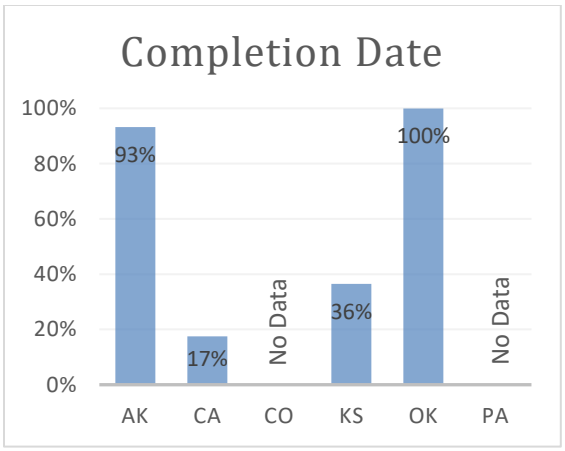
The appendix is provided for a better understand how the total number of wells and fields varies by state.

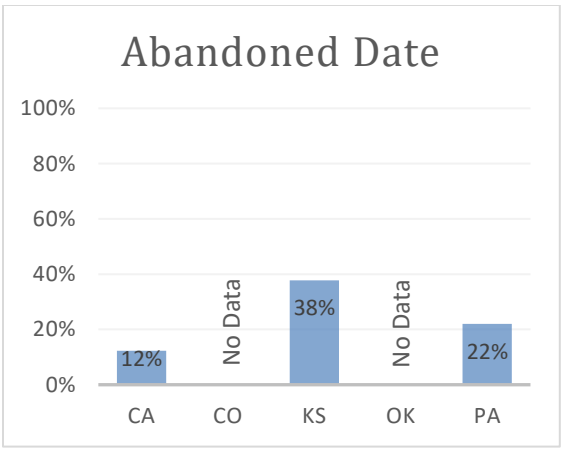
There is a large difference in percent complete of spud date, completion date, and abandon date fields. Furthermore, completion date, abandon date, and depth are reported by only three states. Recall that multiple depth fields are reported, as stated in the Data section. However, there is little standardization as to which types of depths were recorded.

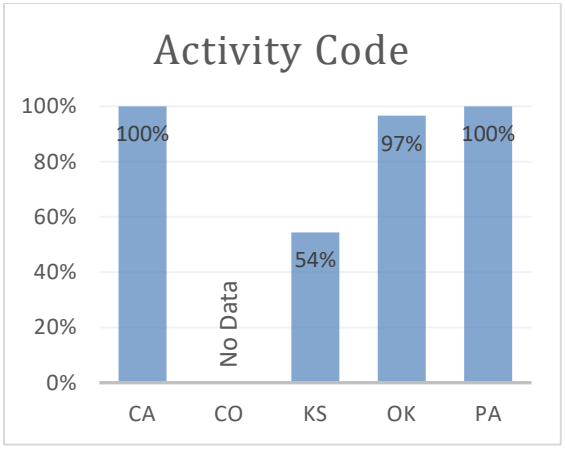
Oklahoma clearly contains the most relevant data out of the states in question while Colorado contains the least data, closely followed by Alaska and Pennsylvania. Most states contain around 90% of the data, barring certain exceptions: California contains 25% or less data on all date fields, Kansas less than 40%, and Pennsylvania less than 55%.

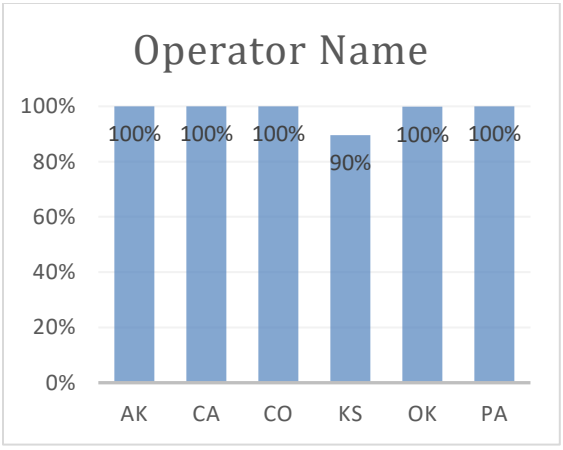

API
100%
80%
60%
40%
20%
0%
100%
100%
100%
97%
100%
100%
AK
CA
CO
KS
OK
PA


Spud Date
100%
80%
60%
40%
20%
0%
93%
25%
70%
34%
100%
53%
AK
CA
CO
KS
OK
PA


Completion Date
100%
80%
60%
40%
20%
0%
93%
17%
No Data
36%
100%
No Data
AK
CA
CO
KS
OK
PA


Abandoned Date
100%
80%
60%
40%
20%
0%
12%
No Data
38%
No Data
22%
CA
CO
KS
OK
PA


Activity Code
100%
80%
60%
40%
20%
0%
100%
No Data
54%
97%
100%
CA
CO
KS
OK
PA


Operator Name
100%
80%
60%
40%
20%
0%
100%
100%
100%
90%
100%
100%
AK
CA
CO
KS
OK
PA

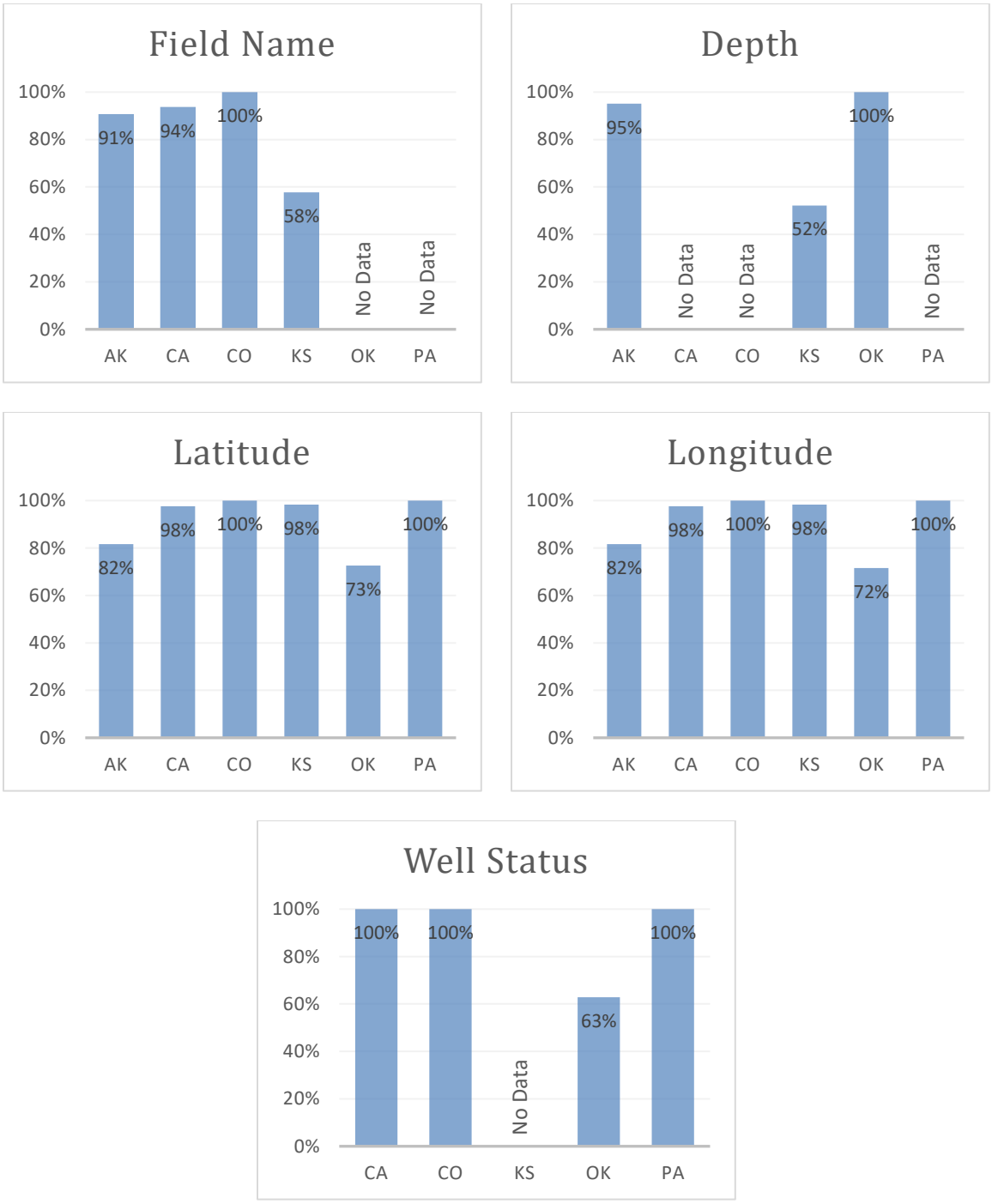
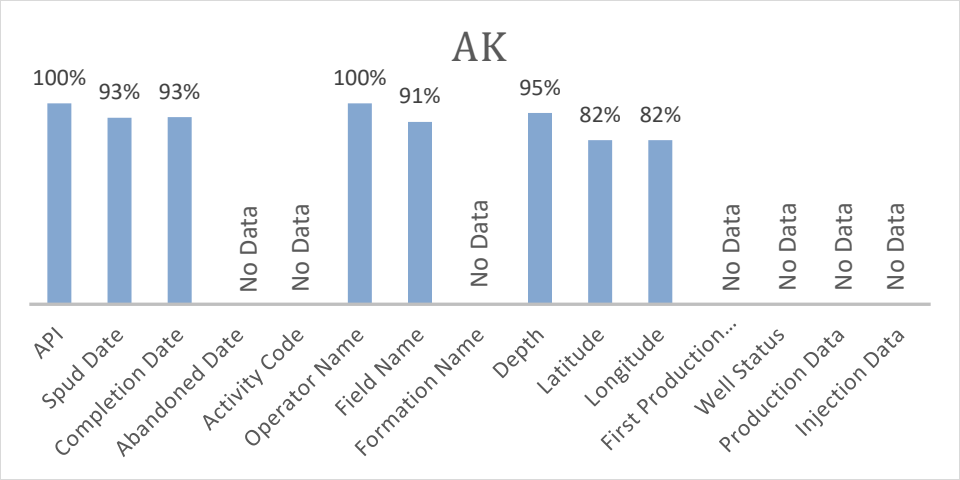


Figure 1: Bar plots displaying percentage of data available for all fields of interest


AK
100%
93%
93%
No Data
No Data
100%
91%
No Data
95%
82%
82%
No Data
No Data
No Data
No Data
API
Spud Date
Completion Date
Abandoned Date
Activity Code
Operator Name
Field Name
Formation Name
Depth
Latitude
Longitude
First Production...
Well Status
Production Data
Injection Data


CA
100%
25%
17%
12%
100%
100%
94%
No Data
No Data
98%
98%
No Data
100%
No Data
No Data
API
Spud Date
Completion Date
Abandoned Date
Activity Code
Operator Name
Field Name
Formation Name
Depth
Latitude
Longitude
First Production...
Well Status
Production Data
Injection Data

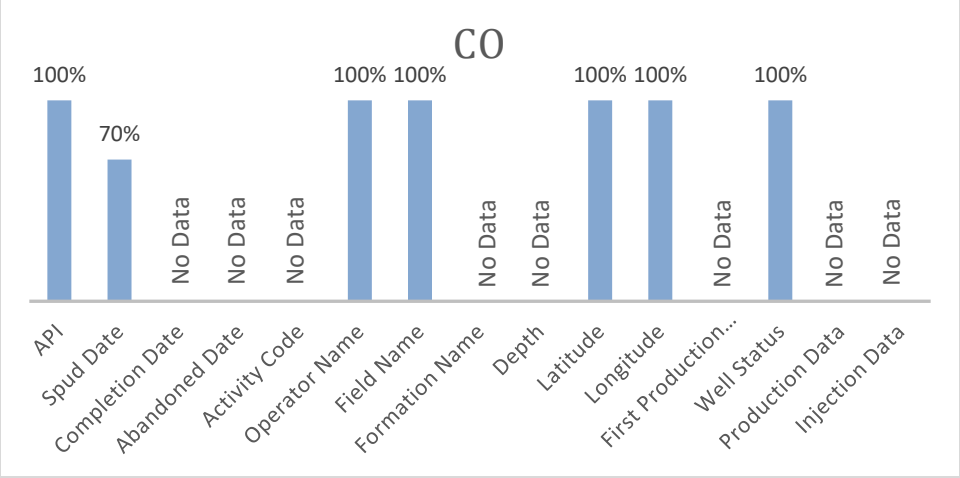

CO
100%
70%
No Data
No Data
No Data
100%
100%
No Data
No Data
100%
100%
No Data
100%
No Data
No Data
API
Spud Date
Completion Date
Abandoned Date
Activity Code
Operator Name
Field Name
Formation Name
Depth
Latitude
Longitude
First Production...
Well Status
Production Data
Injection Data

Figure 2: Bar plots displaying percentage of data available for each state

Graphical

The calendar heat maps for California, Oklahoma, and Pennsylvania appear relatively well distributed—there is no obvious date used as a dummy variable. Alaska and Kansas have a large concentration of dates on January 1, indicating that they either use a date like 1/1/1900 as a dummy variable, or input spuds with an unknown day and month as January 1. Colorado's first column has a similar appearance. This indicates that unknown spud days are relegated to the first of the month they were spudded. Notice that Alaska's legend has a large break; the first four blocks are between 5 and 20 while the last one is around 100—the top left block has a count of 99. Kansas follows a similar trend with most of the counts being less than 700 and only one block being greater than 1,000—the top right block has a count of 1,593.

Colorado and Kansas have one data clock each because they are the only two states whose range of spud dates does not exceed 100 years. Alaska's data clocks appear to be well distributed throughout time, indicating no dummy variables. There is an increase in spud dates from 15 or fewer to between 20 and 40 around 1980, and there is a small concentration of spud dates around January 1905. California is similarly distributed, though it does not appear so unless one looks at the key; there is a jump in number of wells spudded in 2010 by about 100 wells, but besides that there is a natural increase in production. California's dates go back through 1887, which is surprising, but not unreasonable. It is interesting to have data spanning back to the 1800's, though there is a question of increased data uncertainty and record keeping with older wells. Colorado's data clock again shows no unusual trends; there is a jump in the number of wells spudded around 1950 from 30 to 80. The wells spudded in Kansas steadily increases up through 1985, where the number of wells spudded begins to decrease. Oklahoma appears to have a small concentration around January 1900, but besides this, the dates are well distributed until 2010 when there is a spike in number of wells spudded. The dates go through 2028. From 1975 to 1985 and 2000 to 2010 drilling increases in Pennsylvania.

Alaska's line chart shows relatively flat growth with small spikes in 1903 and in the mid-1940s. Following this there is a period of growth that appears to be declining. California is similar, though there is a sharp increase in growth in the mid 2000's that abruptly decreases. Colorado shows a similar trend, though the increase and decrease in the 2000's is more moderate. Kansas shows an almost immediate increase in drilling and a spike around 1980 which then decreases. Oklahoma is steady up until 2000 when there is a sharp spike and a small decrease. There is a small bump around 1900, potentially indicating dummy variables. Pennsylvania mirrors the trends showed by the data clock in that there are two spikes, one in the 80's and one in the late 2000's.

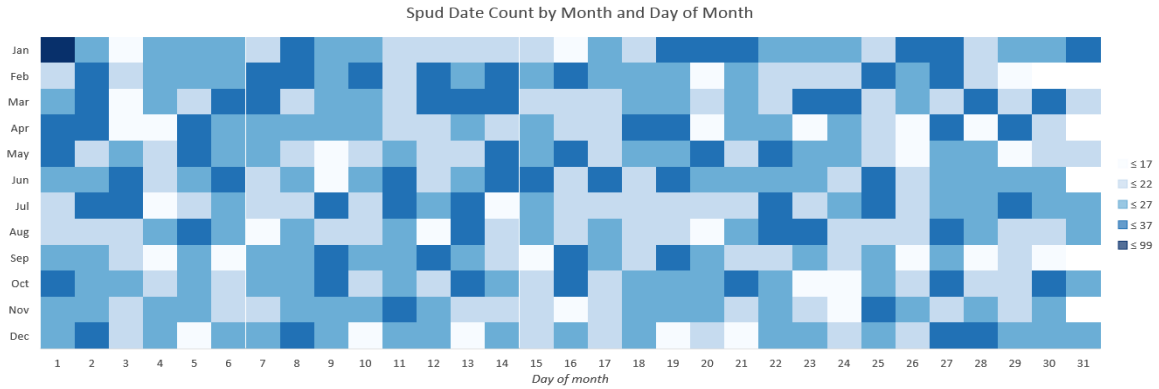


Figure 3: Alaska calendar heat chart

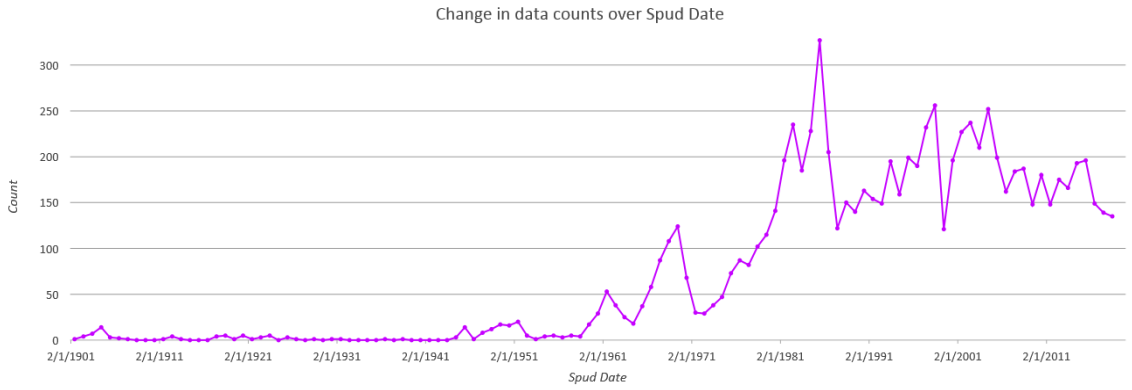


Figure 4: Alaska line chart

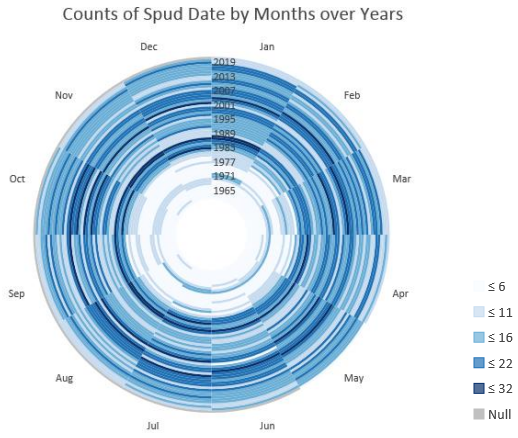


Figure 5: Alaska data clock (1960 to present)

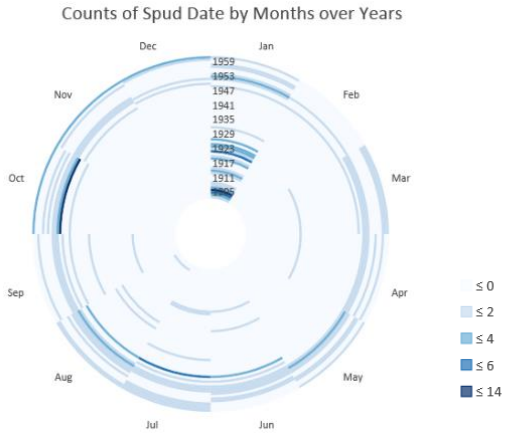


Figure 6: Alaska data clock (pre-1960)

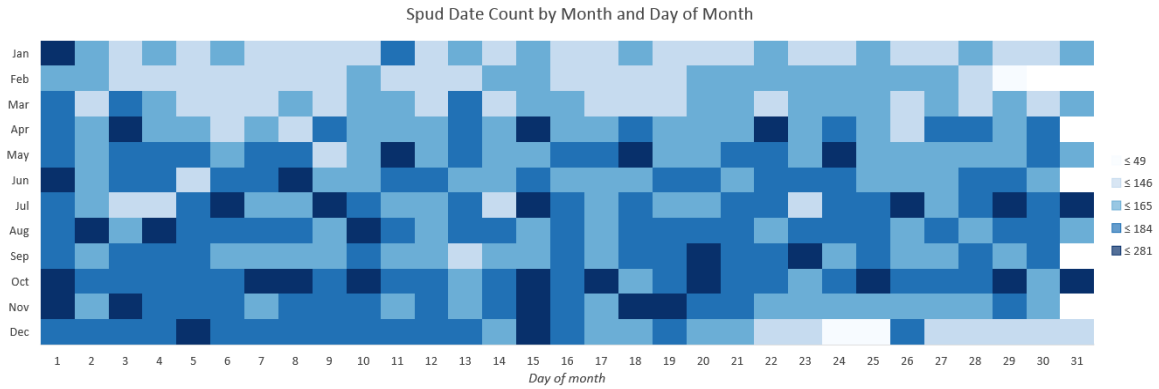


Figure 7: California calendar heat chart

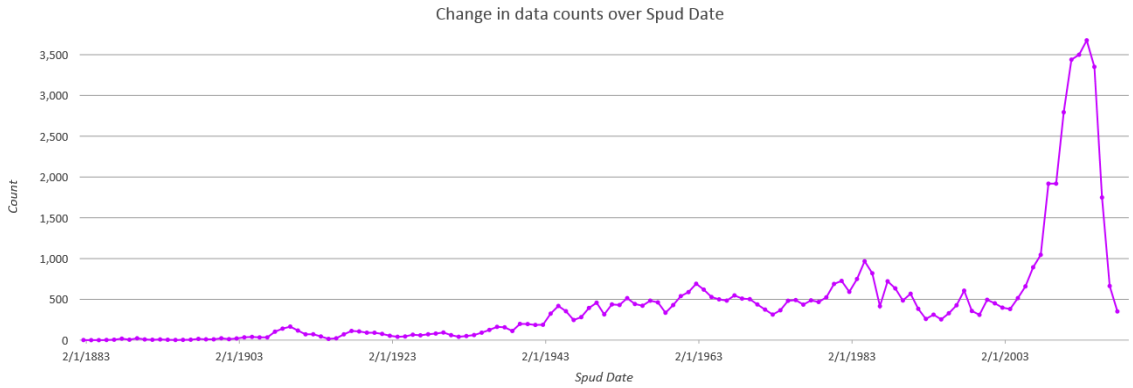


Figure 8: California line chart

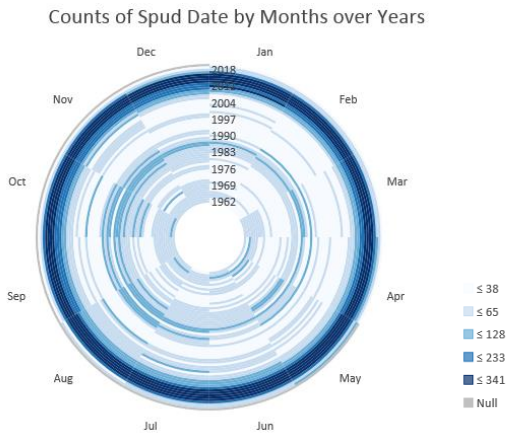


Figure 9: California data clock (1960 to present)

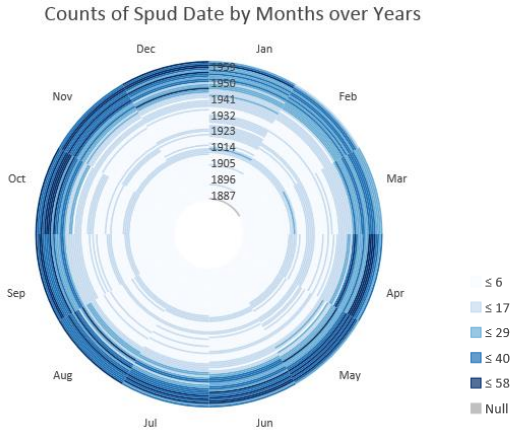


Figure 10: California data clock (pre-1960)

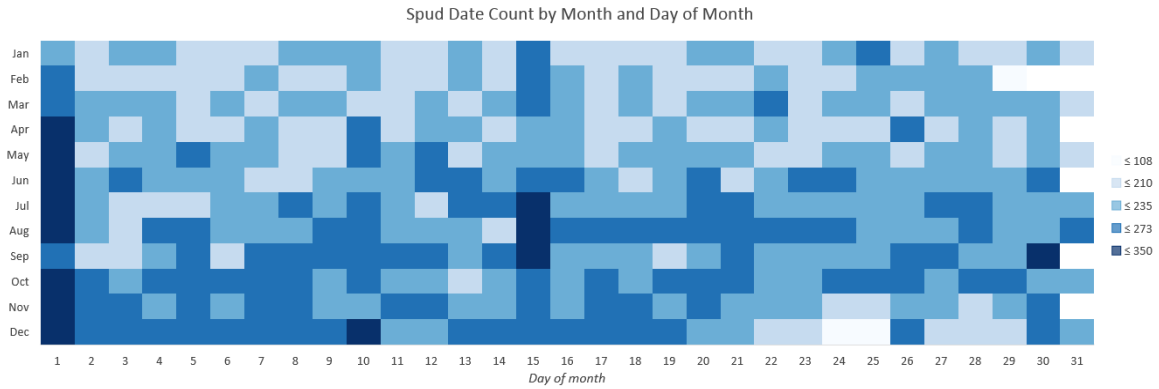


Figure 11: Colorado calendar heat chart

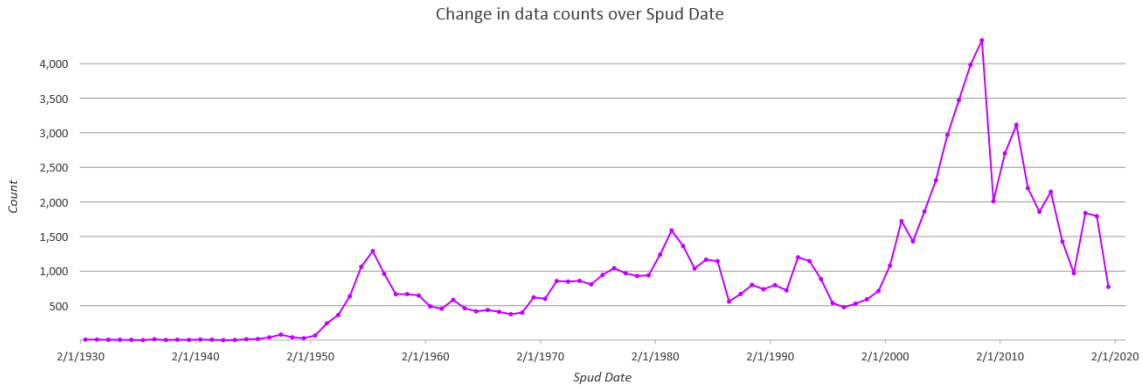


Figure 12: Colorado line chart

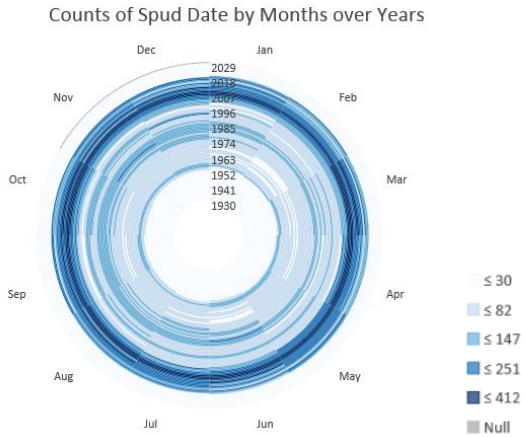


Figure 13: Colorado data clock (1930 to 2029)

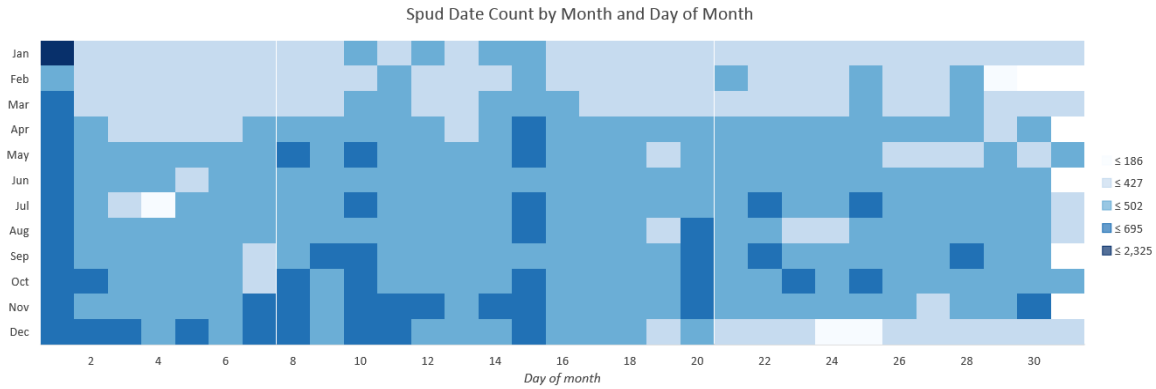


Figure 14: Kansas calendar heat chart

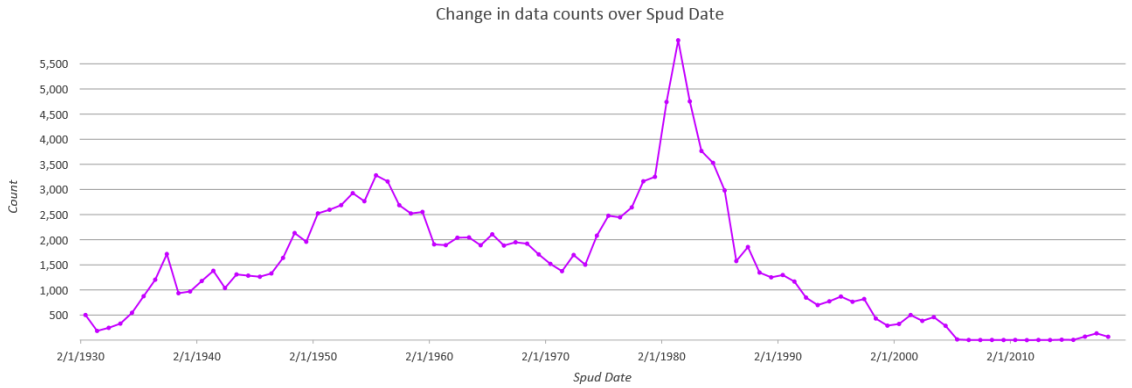


Figure 15: Kansas line chart

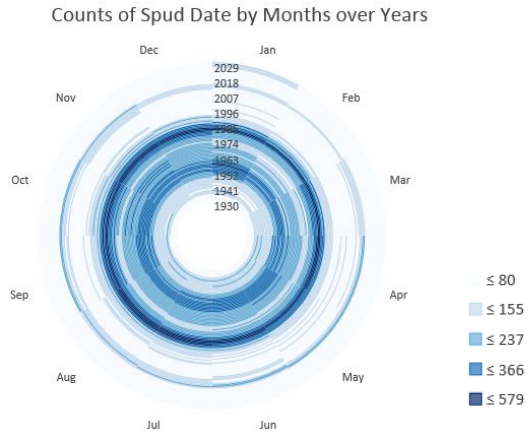


Figure 16: Kansas data clock (1930 to 2029)

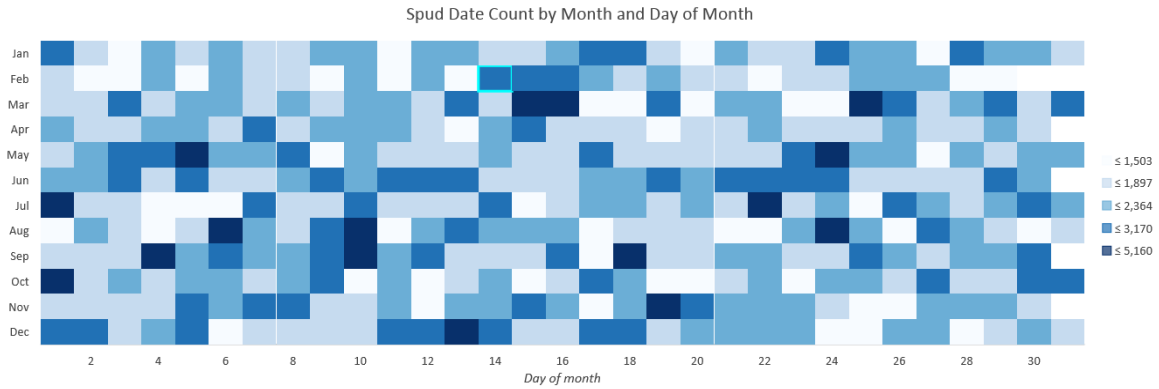


Figure 17: Oklahoma calendar heat chart

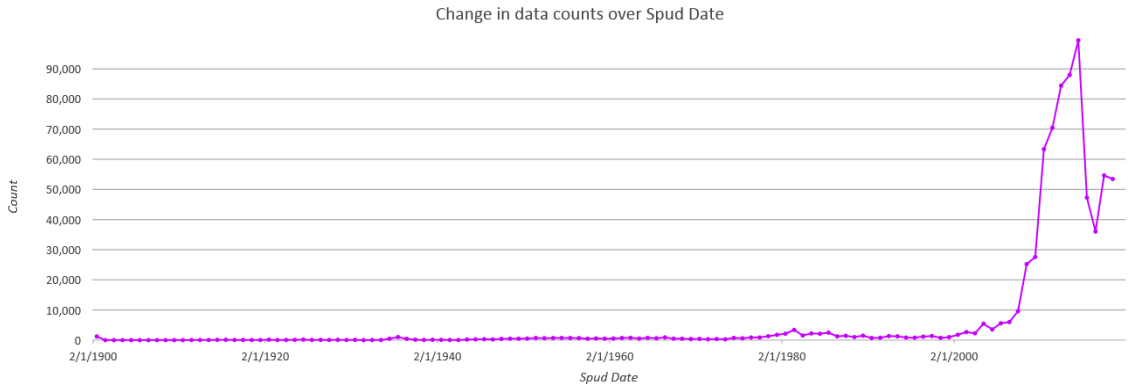


Figure 18: Oklahoma line chart

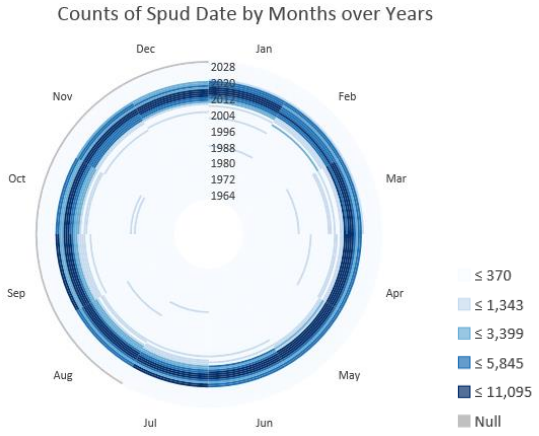


Figure 19: Oklahoma data clock (1960 to present)

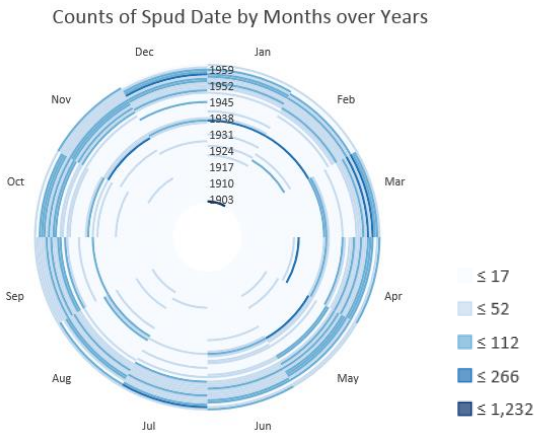


Figure 20: Oklahoma data clock (pre-1960)

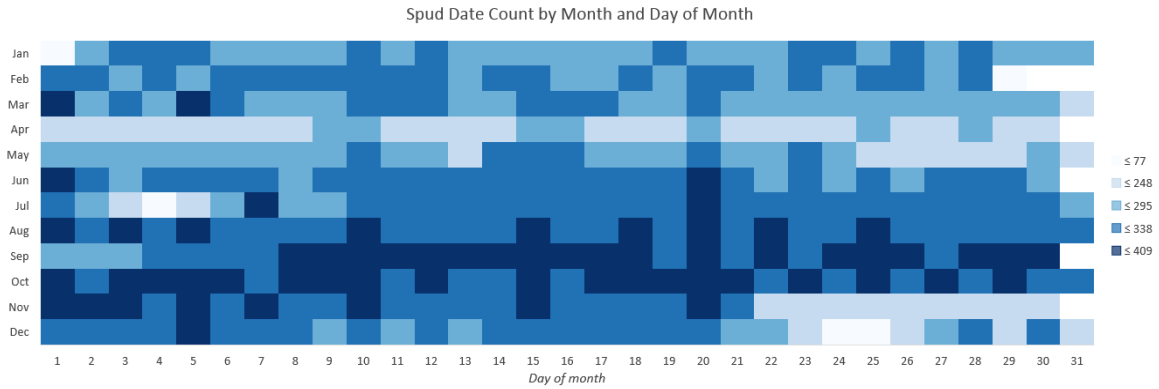


Figure 21: Pennsylvania calendar heat chart

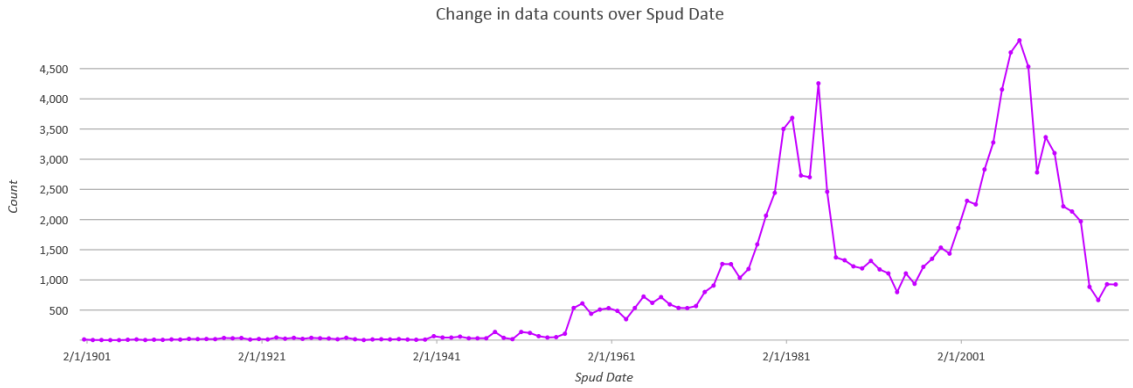


Figure 22: Pennsylvania line chart

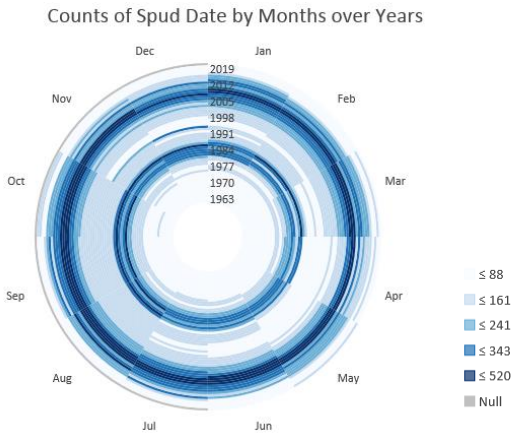


Figure 23: Pennsylvania data clock (1960 to present)

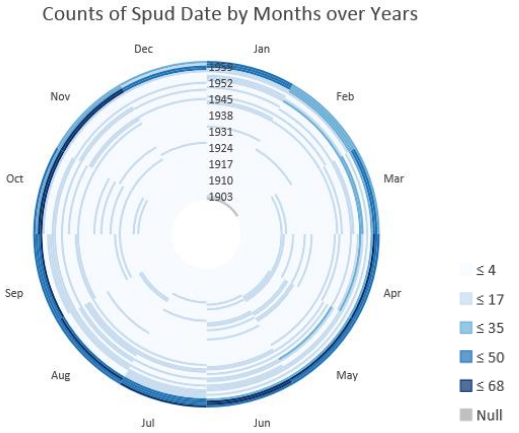


Figure 24: Pennsylvania data clock (pre-1960)

Geospatial

The spud dates are color coded on a gradient from blue to pink, blue being older dates and pink begin more recent dates.

Alaska's wells were initially drilled on the north and south coasts. Later, drilling progressed further out into the ocean. The most recent wells are drilled onshore, again on the north and south coast.

California's wells were initially drilled in central California. They gradually begin to extend further north and south as well as out towards the coast. However, the major concentration of wells remains in central California.

Colorado's wells begin scattered across the state. They gradually concentrated at the northeast corner of the state and at the southwest corner of the state. The northeast pocket extended outward in the central direction. Both clumps of wells then dissipated and scattered before reforming into several small pockets around the corners of the state.

Kansas's wells were initially drilled in central Kansas, extending linearly from the northwest to the southeast corner of the state. This line expanded westward and outward. Another pocket of wells developed in the south west corner of the state which expanded to meet the central line. The explosion of well activity eventually dissipated, leaving only some scattered wells along the southern edge of the state.

Oklahoma began with a large yet scattered number of points. The points eventually converged in the central southern part of the state before scattering again. The points generally remained scattered throughout the state, though the concentration tended to be higher in the center.

Pennsylvania's wells were originally constructed linearly from the southwest corner of the state to the center of the northern boarder. Over time this line extended outward. Other points gradually appeared scattered across the state, though the most concentrated sections remained on that diagonal line and in the northwest corner of the state.

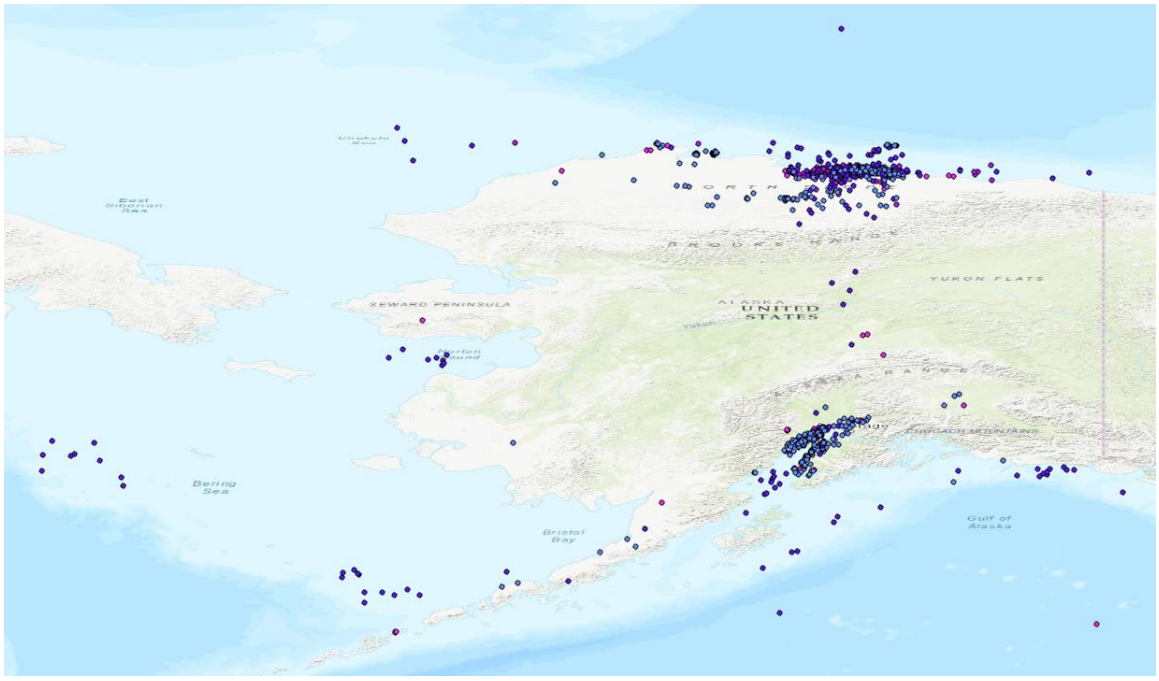


Figure 25: Map of Alaska wells

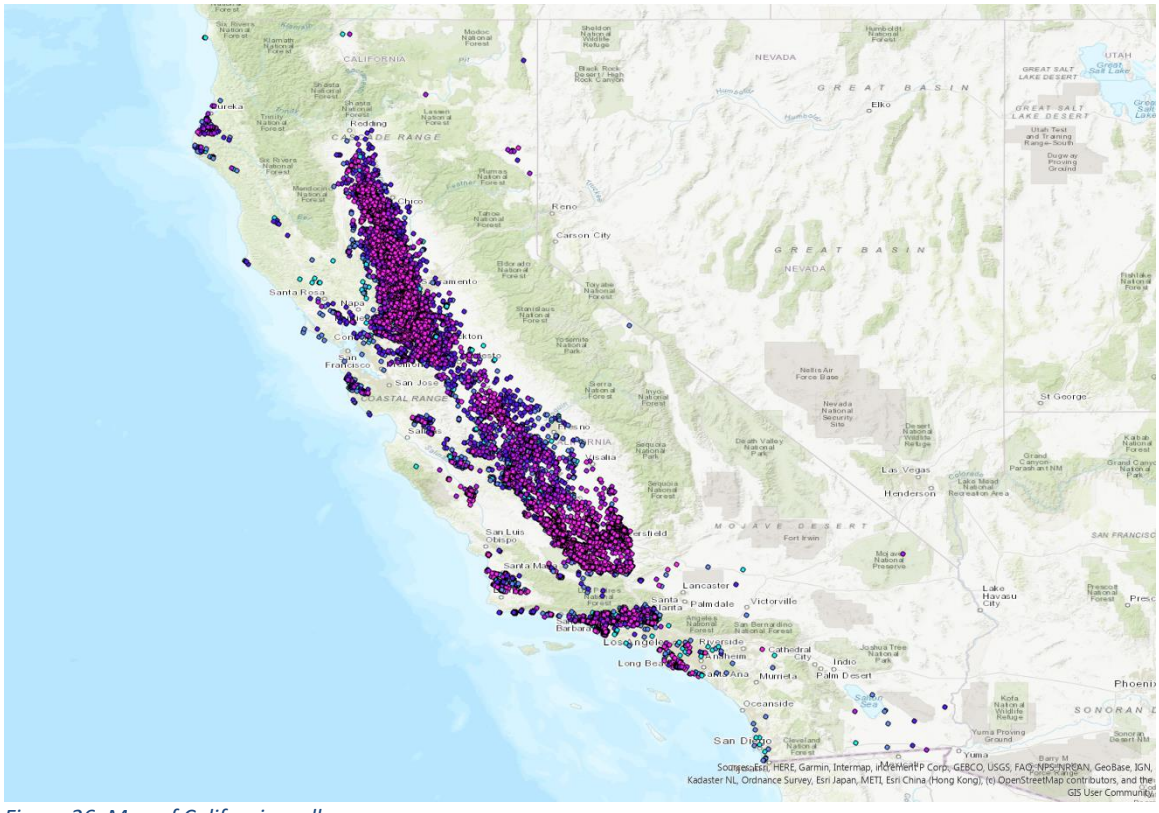


Figure 26: Map of California wells

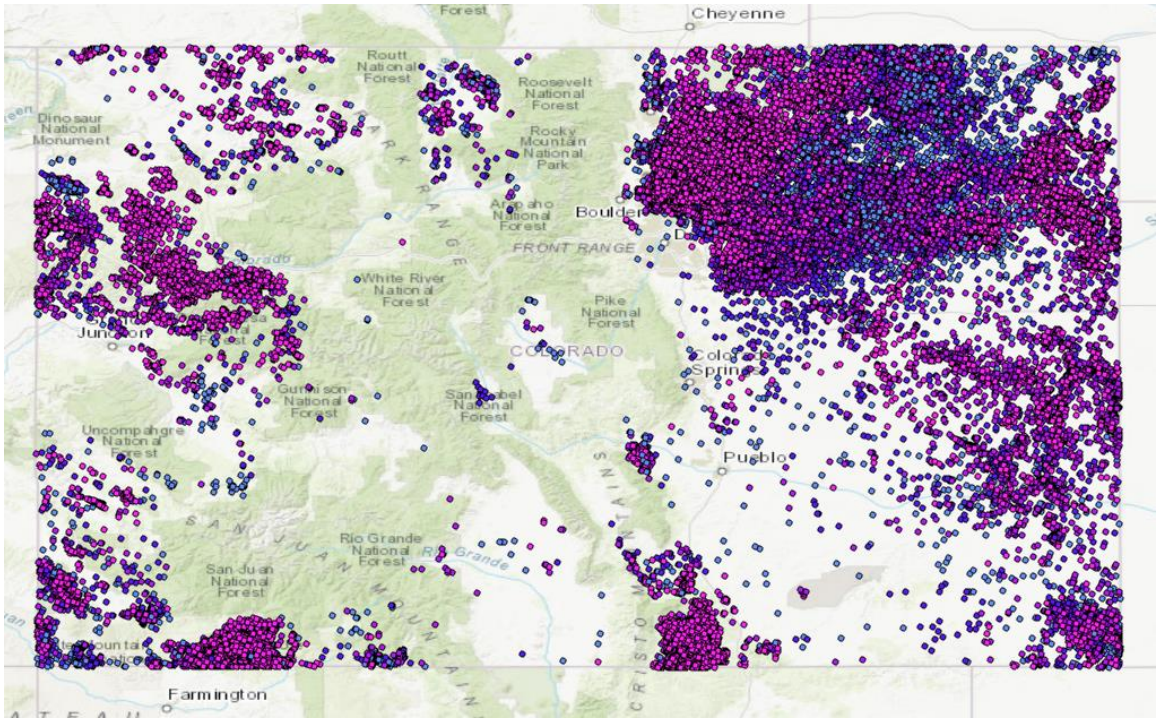


Figure 27: Map of Colorado wells

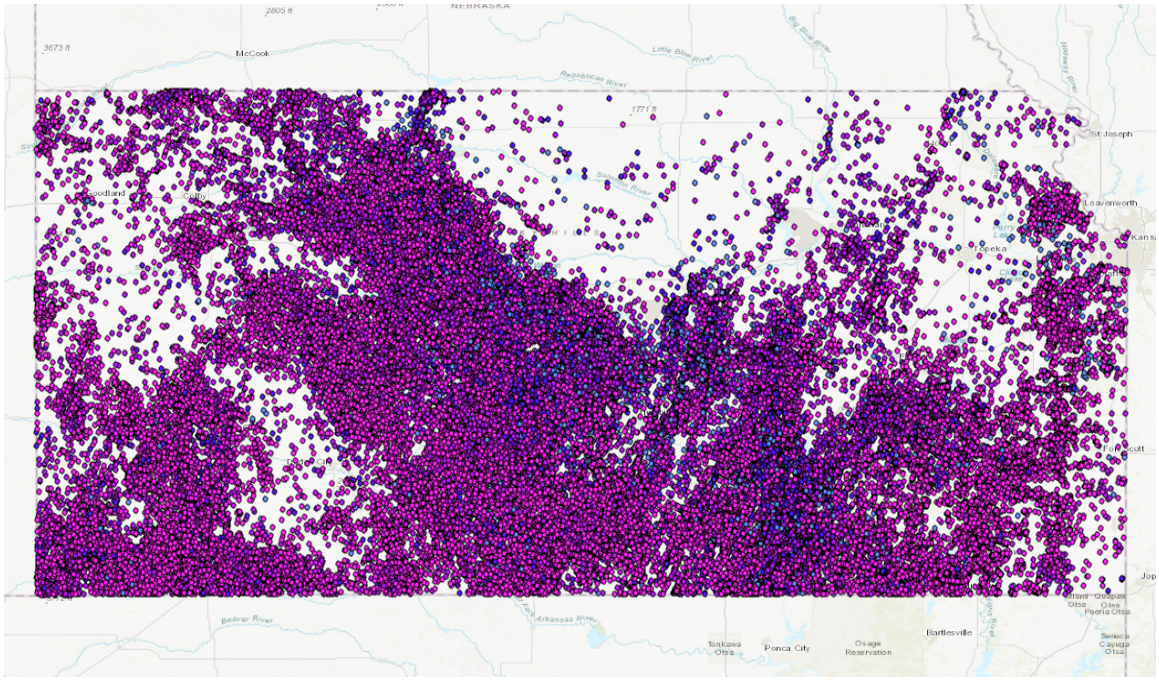


Figure 28: Map of Kansas wells

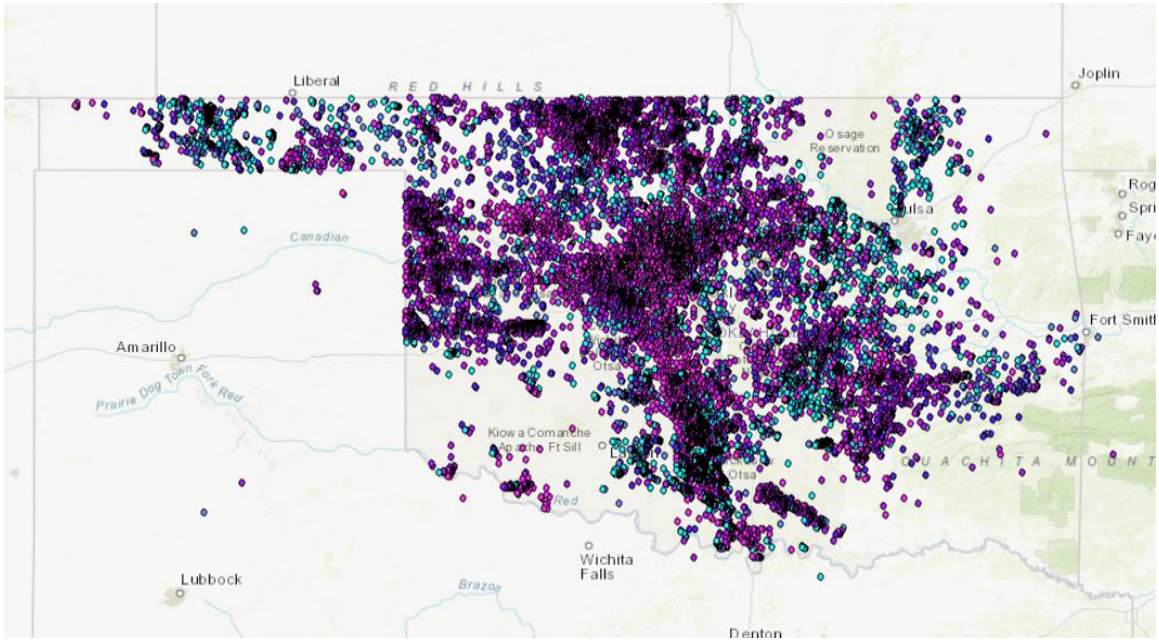


Figure 29: Map of Oklahoma wells

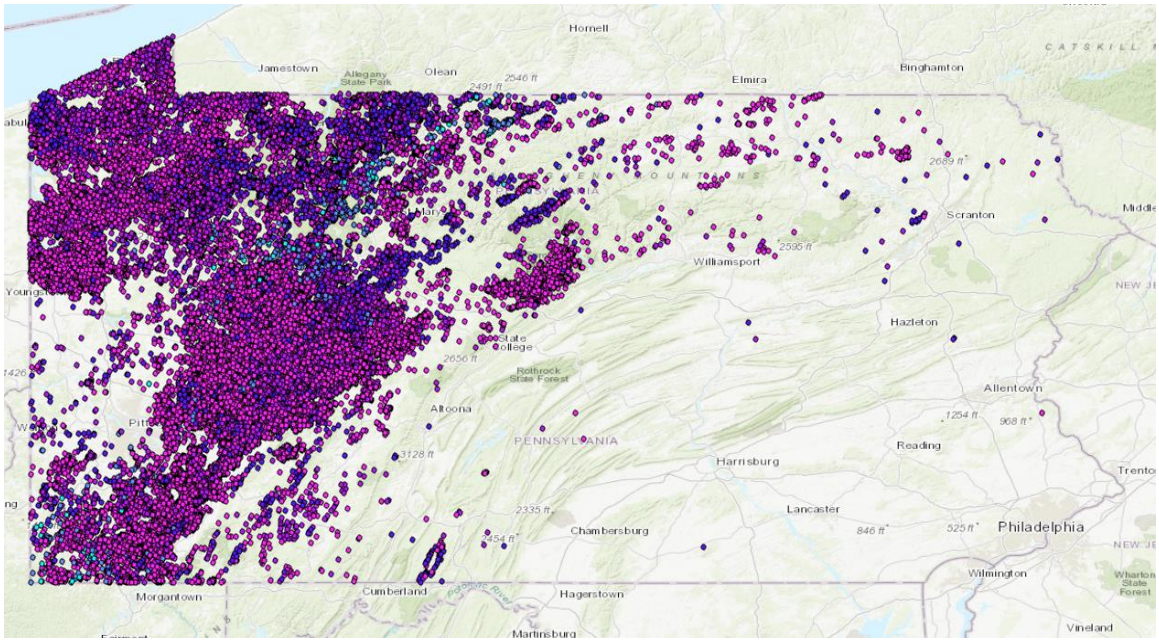


Figure 30: Map of Pennsylvania wells

DISCUSSION

It was expected that Oklahoma would have more data than the other states as it is one of the top oil producers in the country, though the sheer quantity of data available just for Oklahoma was surprising. As the dataset contained many incomplete or erroneous entries, it would be interesting to further investigate Oklahoma. Despite the large quantity of erroneous and dummy variables, the data do not appear to be heavily skewed. Spud date specifically seems to have little effect on the distribution of the calendar heat chart and the data clocks—the data clock shows a larger than normal concentration of data points on January 1900 that is likely due to these dummy variables. However, note that there are similar locations of high concentration amidst lower concentrations throughout the history of Oklahoma's production. More telling is that the calendar heat chart is not affected by the dummy variables; the spread is very even.

Kansas is another state which merits further investigation, though for different reasons. The drilling in Kansas sharply increased in the 1980's and has since decreased. This is due to accelerated gas production from the Hugoton and Panorama fields (18). Wells appear to be distributed over the entire state, so it might be that they have produced as much oil as possible from the state.

Most drilling in Pennsylvania occurs in the Arcadia Basin, also called the Appalachian Forearc Basin. The Marcellus and Utica formations are the two most notable producers of oil and gas. These formations extend up to the northeast corner of the state, so it is surprising not to see a large concentration of wells there. This indicates the dataset is incomplete, or that much of the drilling in that area is confidential. Another indication that the dataset may be incomplete is that Pennsylvania's spud dates only go back to 1900. This might be because wells with those spud dates have not been entered into those datasets, or because those type of wells are filed differently because they were constructed using such different technology.

Other interesting things to note about the spud date were that Colorado, Kansas, and Oklahoma all had spud dates extending out past 2019 to around 2029. These values may represent the expected spud date for permitted, or soon to be permitted wells. The large spikes in drilling around the late 2000's and 2010's shown by the data clocks and line charts are likely due to the fracking boom. Similarly, the increase in drilling in the 1980's in Pennsylvania may be explained by experimental fracking preceding the fracking boom.

CONCLUSION

Knowing the inherent biases and uncertainties in wellbore data is crucial for interpreting results modeling wellbore integrity. Understanding data discrepancies across states can provide insight into influential historic policies and regulation shifts affecting wellbore integrity data trends. This can improve our understanding of wellbore data quality and availability for specific states of interest in the United States. Data deficient and sufficient locations were identified by comparing the data available, ease of access, database structures, and regulatory bodies publishing the data. The findings were synthesized for each location in the form of maps, graphs, preliminary statistical analysis, and written geographic descriptions to provide a comprehensive understanding of wellbore data availability.

Results were coincident with our expectation; there is a large variety of data available and a lot of variation between states. It is interesting to note that the data clock and calendar heat chart were not heavily affected by dummy variables. Most of the data missing from relevant fields were dates—abandoned date, completed date, first production date.

It is likely that the data collected does not contain the total number of wells in each state. This may be because the data is spread across several different datasets or is not entirely controlled by the state—some drilling agencies may be allowed to deny public access to their well logs. Despite the data being incomplete, it still provides a good base for understanding what data is available and how difficult it is to find a comprehensive dataset for each state.

Future Work

Future work on this topic could go in several directions. The most obvious direction is expanding this study to encompass the remaining states; wellbore data could be analyzed via similar methods to compile a comprehensive understanding of publicly available wellbore data in the U.S. This could also be continued at a global level, though there would likely be even larger discrepancies on a country by country basis. This could be studied concurrently with finding complete datasets for each state, possibly in the form of synthesizing multiple datasets from different places together.

Another topic of interest is how publicly available datasets compare to commercial ones like IHS and Drilling Info. This may provide insight into what methods they use to clean the data and how those methods skew the results. This could also inform where else wellbore data may be found, if not from state databases. Lastly, one could study how policy affects wellbore data availability. This would include court cases, laws, and other primary source documents. Understanding the political climate surrounding wellbore data collection can provide historical context for how wellbores are managed and can help evaluate which states are more or less likely to have sufficient, useful, and correct data.

BIBLIOGRAPHY

1. **Dimitroff, Sashe D. and Joy, Michael P.** Oil and gas regulation in the United States. *Thomson Reuters Practical Law.* [Online] June 5, 2019. https://content.next.westlaw.com/Document/I466099551c9011e38578f7ccc38dcbee/View/FullText.html?transitionType=Default&contextData=(sc.Default)&firstPage=true&bhcp=1.

2. **Dunn, Andrea, et al.** *Carbon Storage Atlas (5th Edition).* s.l. : Department of Energy, 2015.

3. **Gerbis, M., Gunter, W. D. and Harwood, J.** Introduction CO2 capture and cological storage in energy and climate policy. [Online] Global CCS Institute, 2019. [Cited: June 3, 2019.] https://hub.globalccsinstitute.com/publications/building-capacity-co2-capture-and-storage-apec-region-training-manual-policy-makers-and-practitioners/introduction-co2-capture-and-geological-storage-energy-and-climate-policy.

4. *"Wellbore Integrity…" Say what??* **Cooper, Julian.** 1, s.l. : Journal Energy Procedia, 2009, Vol. 1.

5. **Carey, Bill.** *Wellbore Integrity and CO2 Sequestration.* Los Alamos : Well Mechanical Integrity Technical Discussion, 2016.

6. *Modeling Gas Miration, Sustained Casing Pressure, and Surface Casing Vent Flow in Onshore Oil and Gas Wells.* **Lackey, G. and Rajaram, H.** s.l. : Water Resources Reserach, 2018, Vol. 55.

7. **Arthur, J. Daniel.** *Understanding and Assessing Well Integrity Relative to Stray Gas Intrusion Issues.* Cleveland : Ground Water Protection Council Stray Gas Forum, 2012.

8. ***Alaska Oil and Gas Conservation Commission.*** **[Online] State of Alaska, 2019. [Cited: June 20, 2019.] http://doa.alaska.gov/ogc/.**

9. *California Department of Conservation.* [Online] State of California, 2019. [Cited: June 11, 2019.] https://www.conservation.ca.gov/dog/Pages/Index.aspx.

10. *Colorado Oila nd Gas Conservation Commission.* [Online] State of Colorado, 2019. [Cited: June 11, 2019.] http://cogcc.state.co.us/#/home.

11. *Kansas Corporation Commission.* [Online] State of Kansas, 2019. [Cited: June 21, 2019.] https://kcc.ks.gov/.

12. *North Dakota Industrial Commission* . [Online] North Dakota Department of Mineral Resources Oil and Gas Division, 2019. [Cited: June 12, 2019.] https://www.dmr.nd.gov/oilgas/.

13. *Oklahoma Corporation Commission* . [Online] State of Oklahoma, 2019. [Cited: June 10, 2019.] http://www.occ.state.ok.us/.

14. *Pennsylvania Department of Environmental Protection.* [Online] Commonwealth of Pennsylvania, 2019. [Cited: June 13, 2019.] https://www.dep.pa.gov/Pages/default.aspx.

15. *Railroad Commission of Texas.* [Online] State of Texas, 2018. [Cited: June 13, 2019.] https://www.rrc.state.tx.us/.

16. Calendar heat chart. *ArcGIS Pro.* [Online] Environmental Systems Research Institute, 2019. [Cited: July 3, 2019.] https://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/charts/calendar-heat-chart.htm.

17. Data clock. *ArcGIS Pro.* [Online] Environmental Systems Research Institute, 2019. [Cited: July 3, 2019.] https://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/charts/data-clock.htm.

18. Carr, Timothy R. *2001 Kansas Oil & Gas Production and Value* . Lawrence : Kansas Geological Survey, 2001.

19. Laumb, Jason. *Summary of Wellbore Integrity Evaluations.* Sacramento : University of North Dakota Energy & Environmental Research Center (EERC), 2016.

20. API Numbering Guidelines. s.l. : API Subcommittee on Well Data Retrieval Systems, 2014.

21. Unique Well Identifier Format. *BC Oil & Gas Commission.* [Online] 2011. [Cited: 7 25, 2011.] https://www.bcogc.ca/sites/default/files/documentation/other-reservoir-engineering/uniquewellidentifierformat_1.pdf.

APPENDIX

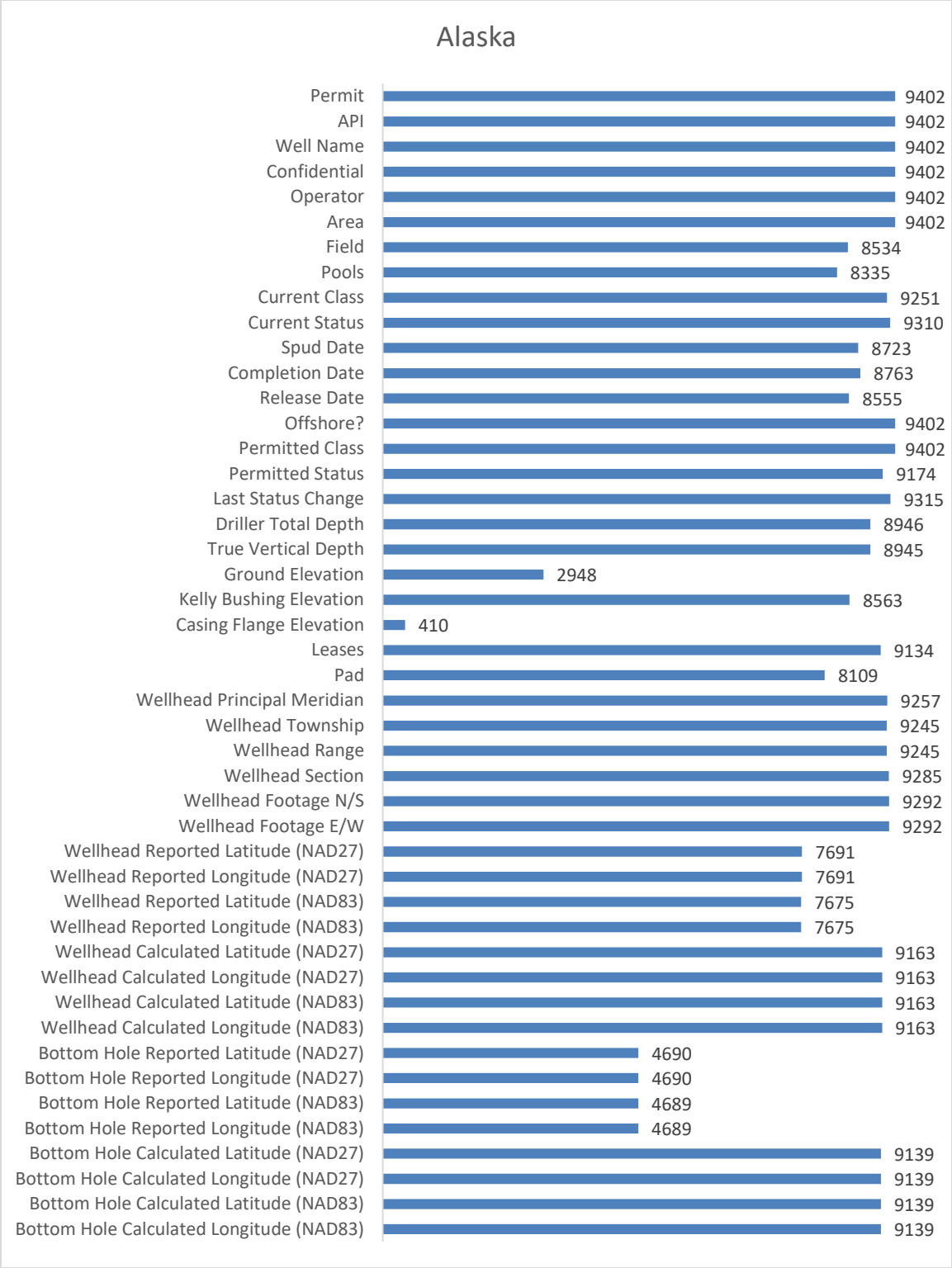


Figure 31: Alaska dataset field counts

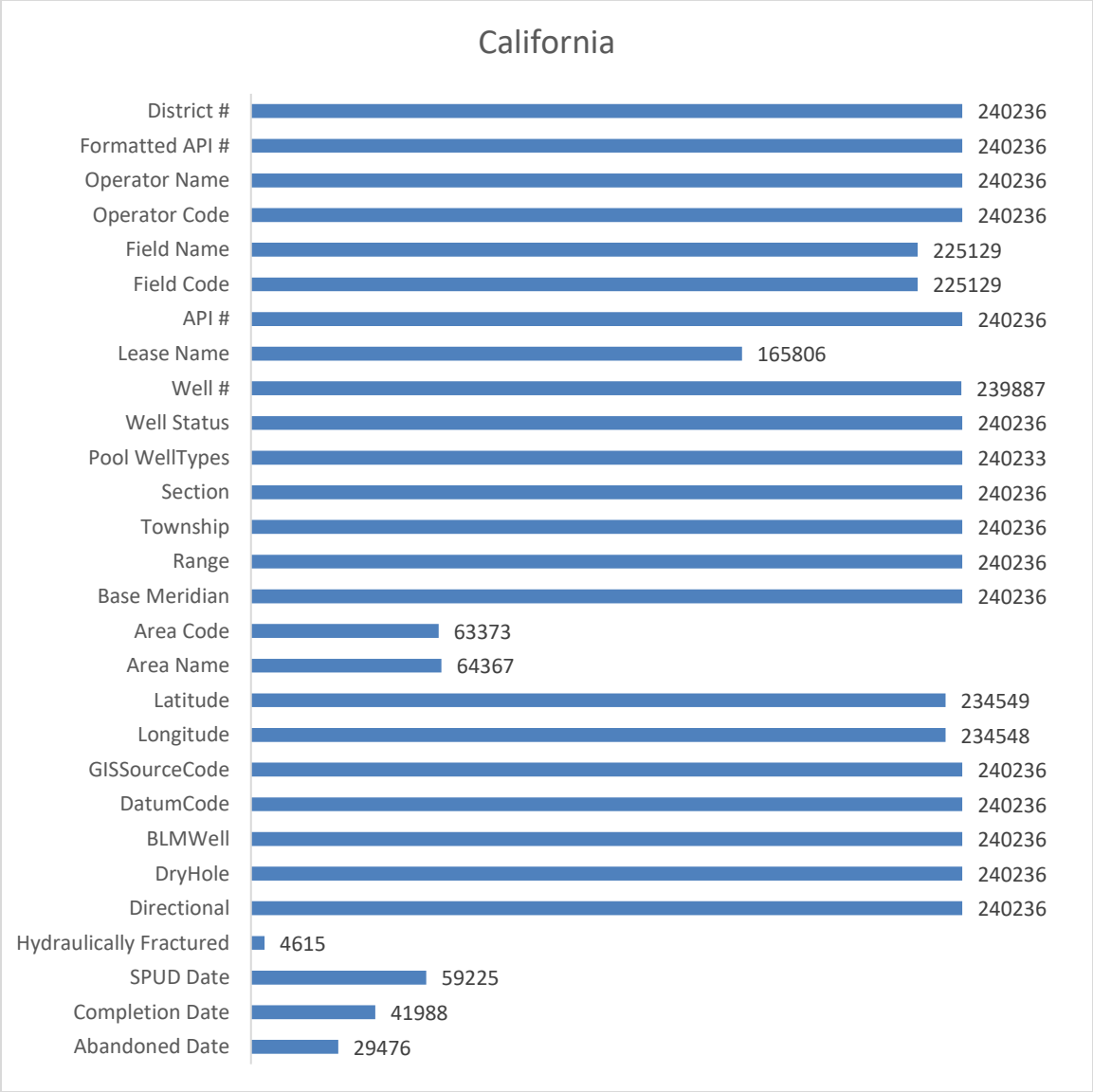


Figure 32: California dataset field counts

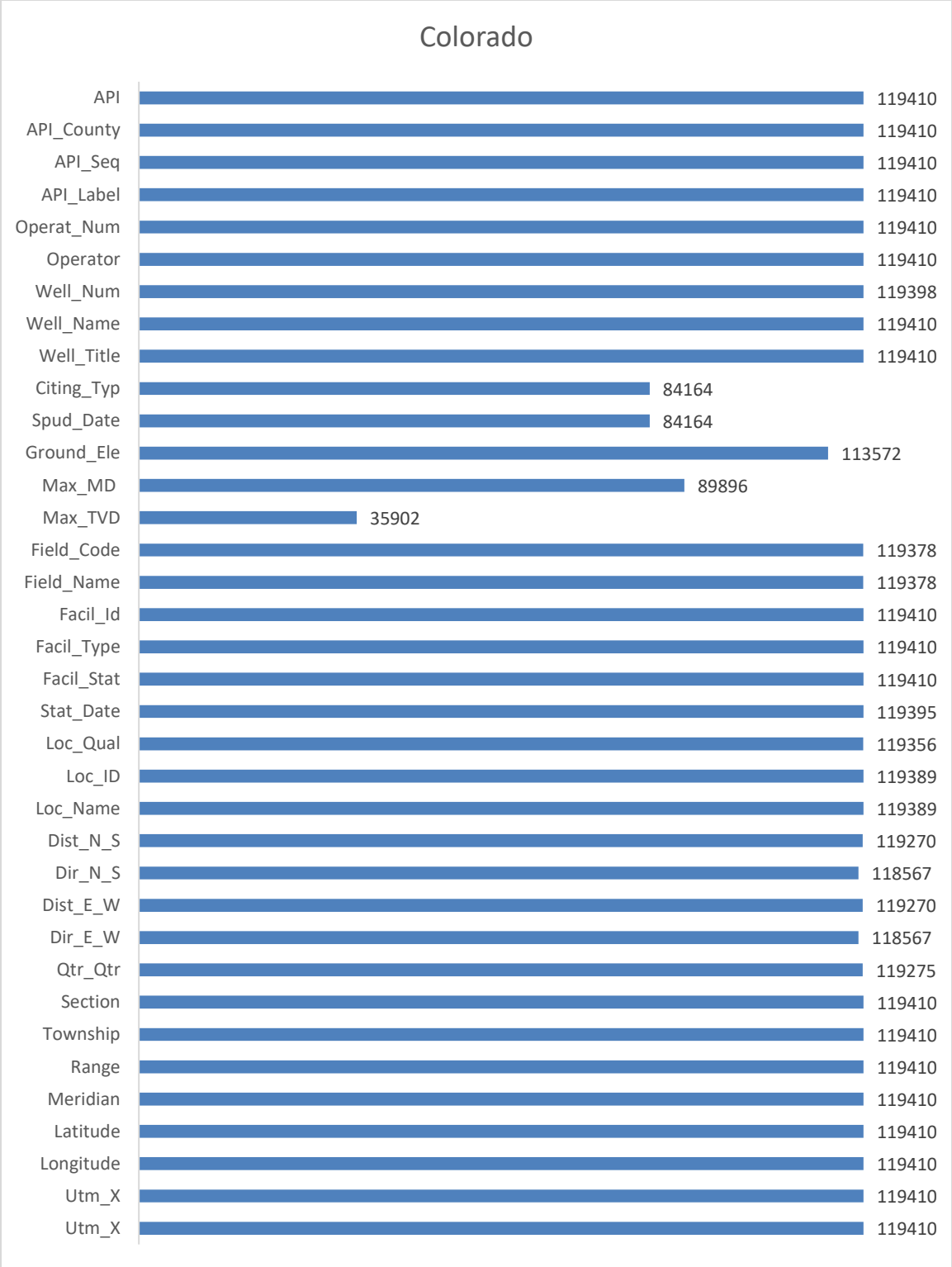


Figure 33: Colorado dataset field counts

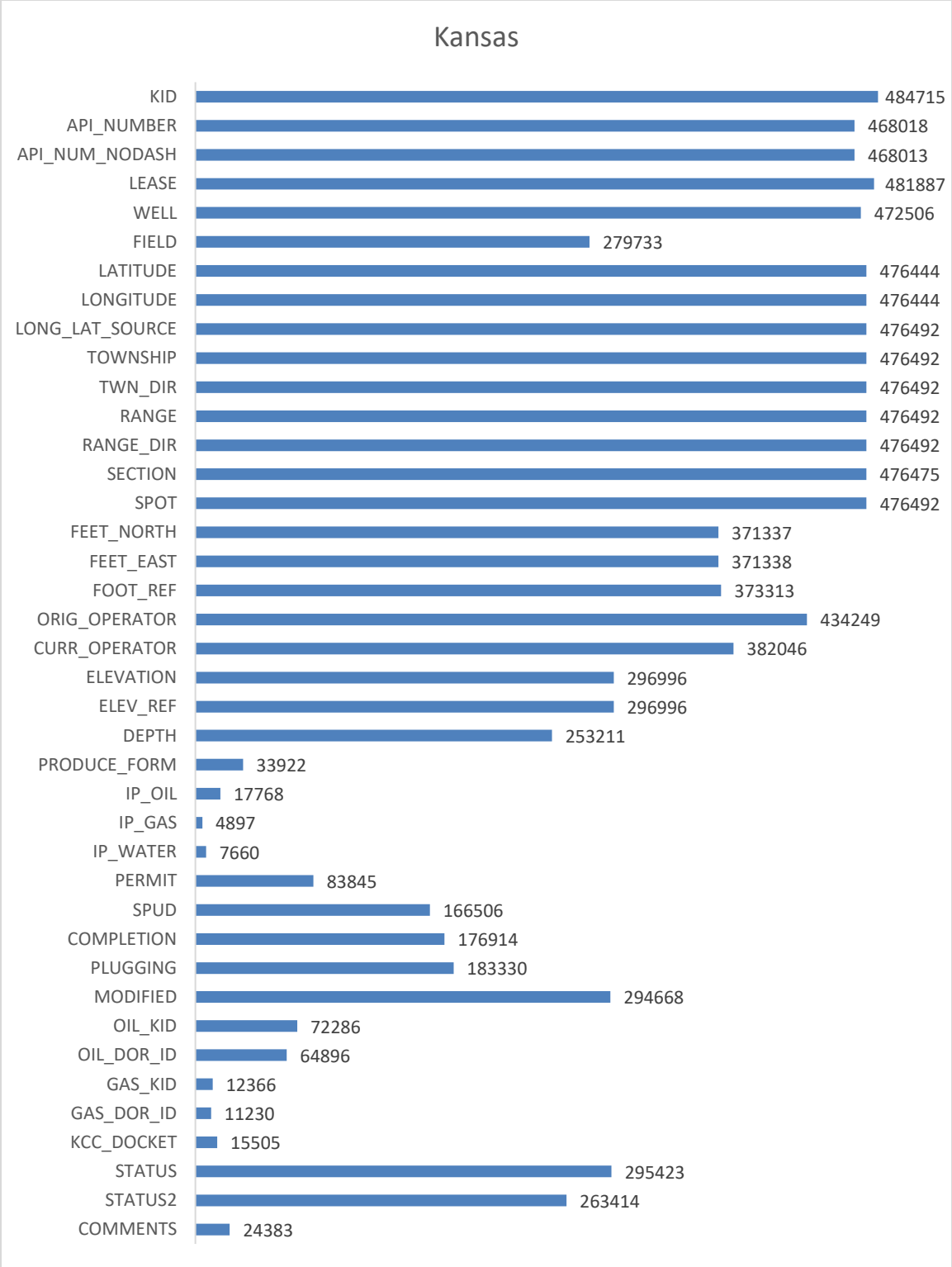
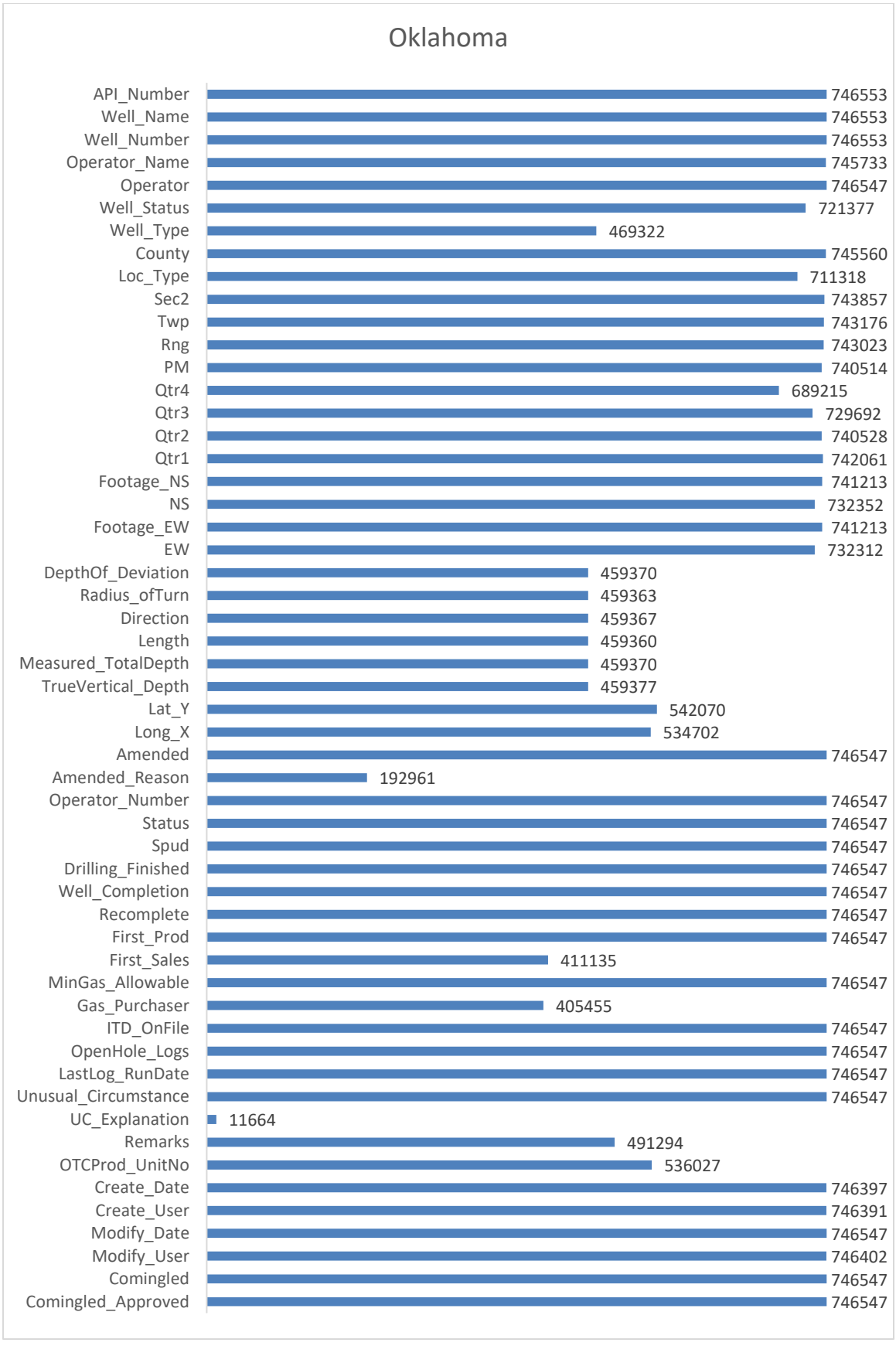


Figure 34: Kansas dataset field counts


Oklahoma
API_Number 746553
Well_Name 746553
Well_Number 746553
Operator_Name 745733
Operator 746547
Well_Status 721377
Well_Type 469322
County 745560
Loc_Type 711318
Sec2 743857
Twp 743176
Rng 743023
PM 740514
Qtr4 689215
Qtr3 729692
Qtr2 740528
Qtr1 742061
Footage_NS 741213
NS 732352
Footage_EW 741213
EW 732312
DepthOf_Deviation 459370
Radius_ofTurn 459363
Direction 459367
Length 459360
Measured_TotalDepth 459370
TrueVertical_Depth 459377
Lat_Y 542070
Long_X 534702
Amended 746547
Amended_Reason 192961
Operator_Number 746547
Status 746547
Spud 746547
Drilling_Finished 746547
Well_Completion 746547
Recomplete 746547
First_Prod 746547
First_Sales 411135
MinGas_Allowable 746547
Gas_Purchaser 405455
ITD_OnFile 746547
OpenHole_Logs 746547
LastLog_RunDate 746547
Unusual_Circumstance 746547
UC_Explanation 11664
Remarks 491294
OTCProd_UnitNo 536027
Create_Date 746397
Create_User 746391
Modify_Date 746547
Modify_User 746402
Comingled 746547
Comingled_Approved 746547

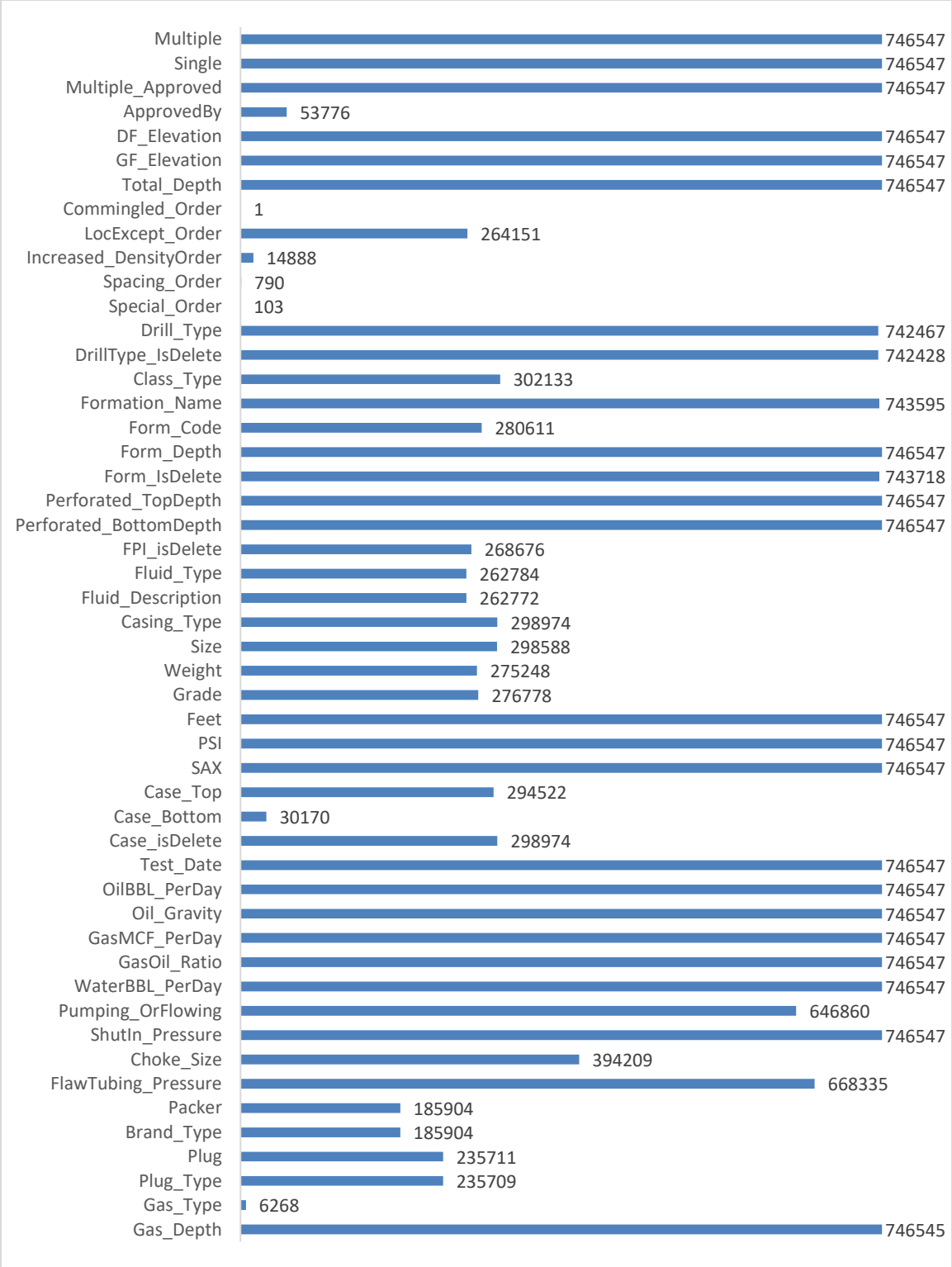


Figure 35: Oklahoma dataset field counts

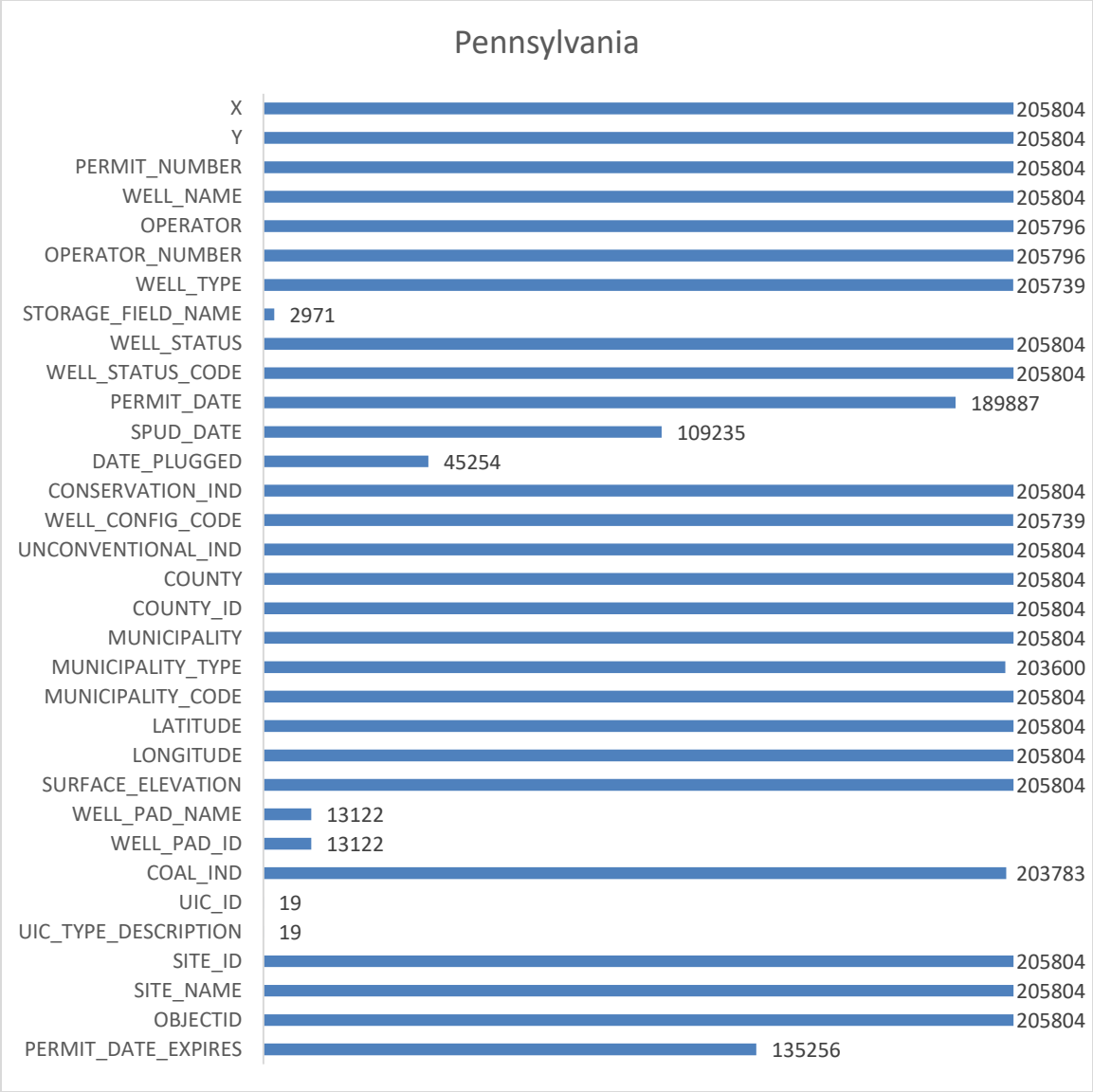


Figure 36: Pennsylvania dataset field counts